Complexification des données et des techniques en linguistique : contributions du TAL aux solutions et aux problèmes

Ludovic Tanguy

Université de Toulouse

Mémoire présenté pour l'obtention d'une
habilitation à diriger des recherches
spécialité : linguistique

Complexification des données et des techniques en linguistique : contributions du TAL aux solutions et aux problèmes

Ludovic Tanguy

CLLE-ERSS : CNRS & Université de Toulouse II-Le Mirail

Complexification des données et des techniques en linguistique : contributions du traitement automatique des langues aux solutions et aux problèmes

Ludovic Tanguy

13 septembre 2012

Table des matières

Sigles et acronymes utilisés

- **ACL** : Association for Computation Linguistics
- **AFC** : Analyse Factorielle des Correspondances
- **ACM** : Analyse des Correspondances Multiples
- **ACP** : Analyse en Composantes Principales
- **ALPAC** : Automatic Language Processing Advisory Committee
- **ANOVA** : Analysis of Variance
- **ATALA** : Association pour le Traitement Automatique des LAngues
- **ATILF** : Analyse et Traitement Informatique de la Langue Française (Nancy)
- **BEA** : Bureau d'Enquêtes et d'Analyses
- **BNC** : British National Corpus
- **CELEX** : Dutch Center for Lexical Information (Nijmegen)
- **CENA** : Centre d'Etudes de la Navigation Aérienne
- **CFH** : Conseil en Facteurs Humains (Toulouse)
- **CHU** : Centre Hospitalier Universitaire
- **CIFRE** : Conventions Industrielles de Formation par la REcherche
- **CLEF** : Cross-Language Evaluation Forum
- **CLEO** : Centre pour l'Edition Electronique Ouverte (Lyon)
- **CLLE** : Cognition, Langues, Langage et Ergonomie (Toulouse)
- **CMLF** : Congrès Mondial de Linguistique Française
- **CoNLL** : Conference on Natural Language Learning
- **CQP** : Corpus Query Processor
- **CRF** : Conditional Random Fields
- **DFKI** : Deutsche Forschungszentrum für Künstliche Intelligenz (Saarbruck)
- **DRT** : Discourse Representation Theory
- **EACL** : European chapter of the Association for Computational Linguistics
- **ECCAIRS** : European Coordination Center for Accident and Incident Reporting Systems
- **ECIL** : Ergonomie Cognitive et Ingénierie Linguistique
- **EI** : Extraction d'Information
- **ELRA** : European Language Ressource Association
- **EMNLP** : Empirical Methods in Natural Language Processing
- **ERSS** : Equipe de Recherche en Syntaxe et Sémantique (Toulouse)
- **ETI** : Ecole de Traduction et d'Interprétation (Genève)
- **GATE** : General Architecture for Text Engineering
- **GREYC** : Groupe de Recherche en Informatique, Image, Automatique et instrumentation de Caen
- **IA** : Intelligence Artificielle

- **IFRI** : Institut Français de Relations Internationales
- **ILF** : Institut de Linguistique Française
- **IHM** : Interface Homme-Machine
- **IMS** : Institut für Maschinelle Sprachverarbeitung (Stuttgart)
- **INIST** : Institut de l'Information Scientifique et Technique (Nancy)
- **INSERM** : Institut National de la Santé Et de la Recherche Médicale
- **IRIT** : Institut de Recherche en Informatique de Toulouse
- **ISSCO** : Institut pour les études sémantiques et cognitives (Genève)
- **KWIC** : KeyWord In Context
- **LERASS** : Laboratoire d'études et de recherches appliquées en sciences sociales (Toulouse)
- **LEREPS** : Laboratoire d'Etude et de Recherche sur l'Economie, les Politiques et les Systèmes sociaux (Toulouse)
- **LIASC** : Laboratoire d'Intelligence Artificielle et Sciences Cognitives (Brest)
- **LISST** : Laboratoire Interdisciplinaire Solidarité, Sociétés, Territoires (Toulouse)
- **LOB** : London/Oslo/Bergen (corpus)
- **LNRE** : Large Number of Rare Events
- **MUC** : Messages Understanding Conference
- **NLTK** : Natural Language Toolkit
- **PAN** : Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection
- **PASTEL** : Programme d'aide à l'analyse sémantique des textes, même littéraires
- **REX** : Retour d'EXpérience
- **RI** : Recherche d'Information
- **RSS** : Really Simple Syndication
- **RST** : Rhetorical Structure Theory
- **SATO** : Système d'Analyse de Textes par Ordinateur
- **SE** : Structure Énumérative
- **SHS** : Sciences Humaines et Sociales
- **SI** : Sémantique Interprétative
- **SIG** : Systèmes d'Information Généralisés (Toulouse, IRIT)
- **SOI** : Sport, Organisations, Identités (Toulouse)
- **STL** : Savoirs, Textes, Langage (Lille)
- **SUR** : Structure à Unité Référentielle
- **SVM** : Support Vector Machine
- **TAL** : Traitement Automatique des Langues
- **TLF(i)** : Trésor de la Langue Française (informatisé)
- **TREC** : Text REtrieval Conference
- **UBO** : Université de Bretagne Occidentale (Brest)
- **UCD** : University College Dublin
- **UMIST** : University of Manchester Institute of Science and Technology (Manchester)
- **UML** : Unified Modeling Language
- **UMR** : Unité Mixte de Recherche
- **UTM** : Université de Toulouse le Mirail
- **XML** : eXtensible Markup Language
- **WAC** : Web As Corpus

Table des figures

Liste des tableaux

Avant-propos

Quand on me demande sur quoi je travaille, je finis généralement par répondre quelque chose du genre « *je m'occupe de corpus et de TAL* ». Dans un cadre plus formel, cela peut prendre la forme nominalisée « *TAL : annotation et exploitation de corpus* », ce qui ressemble plus à un intitulé de cours de sciences du langage qu'à une activité de chercheur. Généralement, je poursuis en précisant « *ces temps-ci je travaille sur X* », *X* étant généralement un type de données, de phénomène langagier ou d'application (les consultations médicales, les citations, les créations lexicales, les structures énumératives, l'attribution d'auteur, etc.), tout en sachant que ce sujet du moment va rapidement être remplacé par un autre.

Quand j'ai fait part à mes proches collègues de cette difficulté à définir plus précisément mon travail, la réponse n'a pas tardé : « *ça, c'est parce que tu n'as pas encore fait ton HDR* ». Dont acte.

Vue d'ensemble

Je vais tâcher de préciser ici le fil directeur des travaux que je vais présenter et problématiser dans ce mémoire.

Si je pense me situer actuellement à mi-chemin entre le traitement automatique des langues et la linguistique outillée, mon parcours a clairement évolué de l'informatique classique vers les sciences du langage. Mes préoccupations principales concernent le développement et surtout l'utilisation de méthodes informatisées pour l'exploitation (au sens large) des données langagières. Une grande partie de mes travaux concernent l'intégration de ces méthodes dans des travaux empiriques d'investigation linguistique. L'autre partie s'adresse plutôt aux côtés applicatifs du TAL, tout en gardant un ancrage fort dans les considérations linguistiques.

Mes préoccupations concernent dans les deux cas l'accès aux données, leur observation et leur analyse. Au fil du temps, elles ont évolué le long d'un axe que l'on pourrait caractériser par une complexification croissante du travail sur les données langagières, dont on peut dégager plusieurs lignes :

- **L'accroissement des volumes manipulés**. Le développement (ou plutôt le retour) de la linguistique empirique a concentré les efforts d'une grande partie de la communauté sur les données langagières numériques. Les corpus sont arrivés au centre de la scène, et ont très rapidement vu leur volume augmenter de façon exponentielle. La dernière évolution en ce sens est bien entendu l'utilisation du Web comme source de données langagières, et un accès à des quantités de textes inimaginables quelques années plus tôt. Mon rôle a été initialement de contribuer à proposer des modes d'accès à ces quantités, que ce soit par le développement d'outils d'interrogation de corpus numériques, ou par la mise en place d'observatoires ciblés du Web. Cette évolution (bien au-delà de

ma contribution) a permis un grand nombre d'avancées positives pour la linguistique descriptive : la principale consiste en une vision plus large, et par là plus précise de phénomènes, voire un renouvellement de certaines questions sur le langage.

- **La complexification de l'outillage**. Les modes d'accès aux données se sont rapidement développés, et leur complexité technique a cru. Les systèmes d'annotation automatique et de fouille dans les textes ont permis un accès plus efficace aux données. Le Web lui-même est devenu un objet encore plus difficile d'accès, au fil de sa propre évolution tant technique que comme objet social. J'ai cherché à prendre en compte cette évolution, en accompagnant les questionnements linguistiques ainsi soulevés, et en proposant leur intégration dans des travaux d'investigation qui les utilisaient. Toutefois, j'ai également constaté un ensemble de problèmes liés à cette technicisation, le principal étant l'éloignement instauré entre la linguistique et son objet d'étude, que les outils informatiques ont participé à creuser.
- **La complexification des données observées**. Parallèlement aux deux points précédents ou à cause de ceux-ci, les données elles-mêmes sont devenues plus riches et par là plus complexes. Les annotations, tant automatiques que manuelles, se sont multipliées et sont devenues le quotidien des études sur les données, rendant les analyses plus ambitieuses, mais aussi plus difficiles. Les besoins en outillage de l'observation sont croissants, mais toujours en retard par rapport aux avancées des volumes et des annotations. J'ai pu proposer à différents moments des méthodes d'observation spécifiques, qui tentaient à la fois de répondre à des questions clairement identifiées, mais aussi d'en soulever de nouvelles. Ces méthodes, ou du moins les besoins ressentis, concernent essentiellement la visualisation des données, seule à même de permettre la compréhension des phénomènes en jeux, et l'émergence d'intuitions nouvelles.
- **La montée en puissance des approches quantitatives**. Plus encore que la linguistique empirique, le TAL a connu et en partie créé ces évolutions. De par son ancrage dans les modèles et les techniques de l'informatique, il s'est donc rapidement tourné vers des méthodes quantitatives, suivant en cela d'autres disciplines que les sciences du langage confrontées, elles aussi, à la montée en volume et en accessibilité des données numériques. À ce stade, le visage du TAL est totalement différent de celui qu'il avait lorsque j'ai commencé mes travaux. Devenues incontournables, les approches basées sur les mesures statistiques et l'apprentissage automatique sont une source d'éloignement culturel dramatique entre la linguistique et le TAL, du moins dans le contexte français. Cet éloignement est d'autant plus regrettable que les articulations sont facilement envisageables, et dans les deux directions. La linguistique empirique peut en effet trouver dans ces nouvelles techniques des outils méthodologiques pour examiner les données volumineuses et complexes, et le TAL peut également intégrer dans ses approches par apprentissage des connaissances linguistiques permettant d'améliorer son efficacité et participer à une meilleure compréhension des phénomènes impliqués. Mon rôle dans cette phase actuelle est de participer à ce rapprochement, en tendant une main de chaque côté.

Quelques précisions préalables

- Je n'ai pas pu éviter la multiplicité des sujets abordés : plutôt que d'en aborder quelques-uns que je considérerais comme centraux, j'ai préféré m'appuyer sur cette diversité pour

faire émerger quelques principes applicables aux nouvelles questions qui ne manqueront pas de se présenter dans l'étude du langage. Ces nouvelles questions, de plus, sont elles-mêmes produites par la mise en relation de sous-disciplines parfois malheureusement cloisonnées, pour lesquelles les approches informatiques et applicatives peuvent être un point de passage.

- J'ai toujours abordé les choses avec un outil (au sens large) issu du monde de l'informatique, que ce soit un programme ou un formalisme de représentation de l'information, de toute façon informatisé lui aussi. C'est dans ce sens-là que je fais du TAL, dans une acception large de la discipline.
- Je n'ai jamais travaillé seul depuis ma thèse : j'ai au contraire passé l'essentiel de mon temps à essayer de comprendre les besoins qui pouvaient s'exprimer autour de moi, et tenter d'y proposer des solutions. Je continue à considérer cet aspect comme une richesse, même si une partie de la communauté académique semble souvent attendre une implication plus personnelle d'un chercheur (et des publications à un seul auteur). Je pense qu'il s'agit là notamment d'une différence culturelle, les disciplines comme l'informatique et les sciences expérimentales fonctionnent différemment des SHS, et le TAL se situe là aussi dans un entre-deux évolutif.
- J'ai toujours « mis le nez » dans des données, et abandonné les approches exclusivement théoriques comme première étape d'une recherche après la fin de mon doctorat. On ne trouvera donc pas, dans mes différentes collaborations, de collaborations ancrées dans une linguistique qui ne serait pas empirique, tout comme je n'ai au final pas cherché à militer pour l'usage des données. J'ai eu la chance de travailler dans un environnement scientifique, et plus précisément à l'ERSS, dans lequel cette conviction est partagée par le plus grand nombre, sinon par tous. Je tiens donc à rendre hommage à ceux qui m'ont précédé et ont su insuffler cette dynamique dans le laboratoire.
- Ma carrière est très influencée par des projets, ce qui n'explique qu'en partie les points précédents (variété des données utilisées, des phénomènes étudiés, des partenariats et des objectifs). Il s'agit là aussi je pense d'une culture initiale venue de ma formation d'ingénieur, confirmée sans doute par l'évolution du contexte plus large de l'administration de la recherche française et de son financement. Si je regrette qu'actuellement la recherche de financement soit imposée à tous les membres de la communauté scientifique, et soit très souvent une perte de temps et une source de frustration, je considère pour ma part que cela m'a souvent permis d'installer des collaborations nombreuses, variées et motivantes.
- Une partie non négligeable de ce que je vais présenter constitue une part souvent cachée d'un travail de recherche, que ce soit les manipulations intermédiaires, les outillages trop classiques pour être détaillés dans des publications, et les tentatives qui n'ont pas abouti à un résultat concluant. Par contre, je pense que ces étapes sont importantes, à la fois comme des exemplifications concrètes des questions plus fondamentales qu'elles soulèvent, et en tant qu'éclairage sur les pratiques en évolution du TAL face aux données et aux questions linguistiques.

Quelques points de vocabulaire

Je tiens à préciser ici, à travers une série de définitions, ce que j'entends par certains des termes centraux utilisés dans mon travail et dans ce mémoire. Ces définitions n'ont pas de

prétention à l'objectivité, mais servent à éclairer ma vision des concepts sous-jacents.

Données langagières

On verra dans la suite que ce que j'entends par *données langagières* recouvre une réalité des plus disparates : des textes bruts (voire très bruts comme les pages Web), des textes annotés globalement (et automatiquement) ou localement (par annotation manuelle ciblée), des informations extraites de textes annotés (segments, structures), des lexiques (mais quand c'est le cas, *via* un retour systématique à la source et à une analyse du contexte), et des choses parfois totalement inclassables.

Traitement automatique des langues (TAL)

Dans l'acception très large que je lui attribue, le traitement automatique des langues recouvre l'ensemble des travaux qui articulent l'outillage informatique et les données du langage. Comme l'équilibre entre l'informatique et la linguistique est instable, ce terme regroupe des approches très diverses, et des visées qui recouvrent aussi bien le développement d'applications que l'outillage de l'investigation de données à des fins de description linguistique.

En tant que discipline, le TAL regroupe un ensemble de pratiques qui forment son noyau central. Dans celui-ci on retrouve les procédures automatisées de projection d'information sur des données (textes, lexiques, etc.), et leur exploitation par des approches calculatoires.

Linguistique de corpus

Au sein de la linguistique, les approches sur corpus forment une famille difficile à délimiter, dont les liens avec le TAL sont souvent très resserrés. Dans ma vision actuelle, le cœur de la linguistique de corpus consiste en un ensemble de techniques et de méthodes d'investigation de données rassemblées dans un but particulier de description des mécanismes du langage. Que le corpus soit vu comme un échantillon représentatif d'un ensemble plus large, ou comme un objet fini délimité par des caractéristiques extra-linguistiques (liées aux conditions de leur production), la linguistique de corpus met en œuvre des approches informatisées pour tester des hypothèses ou mettre au jour des régularités nouvelles.

La question de la quantité de données manipulées, et l'évolution de cette quantité au fil des progrès technologiques est centrale. La mesure de la fréquence des observables, et l'interprétation des valeurs, en regard ou non d'autres paramètres est un instrument incontournable de toute approche sur corpus. De ce fait, les méthodes quantitatives (dont je détaillerai certaines) y sont nécessaires, et leur informatisation croissante est un des enjeux qui font de la linguistique de corpus un important consommateur des développements du TAL.

Recherche d'information (RI)

En tant que discipline de l'informatique, la recherche d'information vise à permettre et faciliter l'accès à une sous-partie pertinente d'une collection de données. La recherche d'information textuelle (la seule avec laquelle j'ai été en contact) est donc très intimement liée avec le traitement automatique des langues, puisqu'elle cherche à représenter les documents d'une collection de telle façon qu'ils puissent être mis en regard avec l'expression d'un besoin d'information. Le développement de la RI a produit des techniques capables de permettre un accès à des quantités croissantes de données.

On verra donc que les points de contacts de la RI textuelle avec mes préoccupations sont de deux types : exploiter les techniques de RI pour accéder aux données langagières, notamment lorsqu'elles sont volumineuses, et voir en quoi les modes de représentation des données du langage peuvent être intégrés dans les processus de RI eux-mêmes.

Extraction d'information (EI)

Souvent présentée comme une sous-discipline de la recherche d'information, l'extraction d'information vise à l'identification d'éléments informatifs dans une collection de documents. Elle propose ainsi une série de techniques permettant de structurer l'information contenue dans des textes, en recherchant des configurations langagières particulières. Il est donc plus naturel de voir l'EI comme une famille des approches du TAL, en ce sens qu'elle développe des techniques automatiques d'analyse des données langagières. Si ces techniques ont généralement des visées applicatives (extraction de connaissances), elles entretiennent le même double rôle que celles de RI avec l'investigation des données en linguistique : elles permettent de repérer des structures particulières dans les textes qui font partie des objets d'étude de la linguistique de corpus et elles peuvent bénéficier des connaissances et des ressources de la linguistique pour ce faire.

Malgré la grande proximité de certains de mes travaux avec l'EI, j'ai encore un peu de difficulté à les rattacher à cette discipline, dont les frontières me semblent difficiles à cerner ; ceci explique notamment la rareté des mes emplois de ce terme.

Méthodes quantitatives

Je regroupe sous ce terme générique un ensemble très disparates de méthodes d'analyse des données (langagières). Comme je l'ai dit, l'étude de la fréquence et de la distribution des observables en linguistique de corpus fait appel à des analyses et des modèles statistiques, qui constituent donc les méthodes quantitatives centrales. A celles-ci j'ajoute de façon plus périphérique les modes de représentation graphiques et de visualisation de l'information, qui comme les approches statistiques permettent une vision synthétique d'une collection de données. Mais le point focal de cette dénomination regroupe les techniques plus récentes et situées actuellement sur le devant de la scène du TAL que sont les techniques d'apprentissage automatique et de fouille de données.

Apprentissage automatique

L'apprentissage automatique (ou *artificiel* et par la suite *apprentissage* tout court) est une discipline issue de l'intelligence artificielle qui vise à construire des outils informatiques capables de traiter des données sans que la procédure de traitement ait été totalement détaillée. Ces méthodes se basent, en lieu et place de l'explicitation des calculs à appliquer, sur l'observation d'un ensemble de données en faisant émerger des régularités qui sont ensuite formalisées et exploitées. L'approche prototypique de l'apprentissage automatique (dit *supervisé*) consiste à fournir au système un ensemble de données pour lesquelles la réponse souhaitée est connue, pour qu'il construise sur la base d'analyses statistiques un modèle, reliant les caractéristiques descriptives des données à la réponse visée. Une fois le modèle obtenu, le même système est capable de l'appliquer à de nouvelles données similaires à celles sur lesquelles il s'est basé pour produire une réponse. Le travail de développement est donc grandement simplifié, puisqu'il ne nécessite pas (ou peu) d'explicitation des connaissances formalisées du phénomène visé.

L'utilisation des techniques d'apprentissage automatique en TAL a cru exponentiellement depuis plus d'une vingtaine d'années, si bien qu'elles constituent actuellement le mode principal de réponse aux besoins applicatifs, en lieu et place des modèles formels du langage.

Fouille de données

La fouille de données recouvre un ensemble de techniques d'investigation des données qui utilisent les mêmes principes que l'apprentissage automatique. La différence principale entre les deux familles concernent leur objectif : là où l'apprentissage automatique vise au développement d'un système apportant une caractérisation des données, la fouille de données est concernée par l'extraction de connaissances nouvelles. En se basant de la même façon sur une collection de données, les méthodes de fouille utilisent des méthodes quantitatives pour repérer des régularités, qui peuvent par exemple correspondre à des relations entre certaines caractéristiques des données, ou à l'identification de classes au sein de la collection.

La fouille de données langagières (également appelée fouille de textes) rejoint donc les préoccupations de l'extraction d'information, mais en faisant généralement appel à des techniques plus massives.

Plan du mémoire

La présentation est guidée par les types d'approche et de méthodes que j'ai appliquées ou contribué à mettre au point :

- Partie 1 : présentation chronologique de mon parcours, en identifiant les différentes rencontres et les influences qu'elles ont eues sur mon travail. Dans cette partie j'essaie également d'identifier les principales évolutions plus larges qu'ont vécues les sciences du langage et le TAL.
- Partie 2 : outils d'accès au contenu textuel - recherche dans les textes et utilisation du Web comme corpus. Cette partie traite donc la complexification du travail du linguiste par le fait qu'il est confronté à des données plus volumineuses, accessibles uniquement par le biais d'outils d'interrogation, en dégageant les points positifs et négatifs de cette évolution.
- Partie 3 : méthodes de représentation des données et analyses statistiques. J'y montre comment l'accroissement des volumes et la multiplicité des annotations crée de nouveaux besoins en termes d'examen des données langagières, et les différentes méthodes envisagées pour les observer.
- Partie 4 : fouille de données (pour explorer des annotations complexes) et apprentissage automatique (pour répondre à des besoins précis concernant les données langagières). J'aborde un ensemble de questions liées à l'utilisation de ces techniques pour un travail de description linguistique, mais surtout comment ces techniques sont un lieu plus difficile de collaboration entre la linguistique et le versant plus technique du TAL tel qu'il s'est développé depuis quelques années.

J'ai placé à différents endroits du mémoire des frises chronologiques qui tentent de donner une vue d'ensemble de l'évolution de certaines questions que je traite. À chaque fois, j'ai choisi de mettre en évidence certains points et événements afin de les mettre en perspective. Ces frises sont donc très nettement orientées, et n'ont pas la prétention d'être exhaustives ni parfaitement objectives. L'établissement de ces outils argumentatifs et d'aide à la lecture correspond à un besoin, dans cette phase de ma carrière qui correspond à un bilan d'étape,

de réfléchir à ma position dans le paysage de ma discipline, bilan qui ne peut s'abstraire d'une dimension temporelle. Je tiens également à préciser ici qu'il n'est pas toujours aisé de retrouver les informations précises concernant les avancées techniques et scientifiques un peu plus reculées (disons, datant d'avant l'arrivée du Web). Outre le fait que certaines imprécisions ou erreurs sont inévitables, espérons que cela suscitera à mes lecteurs l'envie de les corriger et de les compléter. Espérons aussi que certains événements, comme le cinquantenaire de l'*Association for Computational Linguistics* qui aura lieu à l'automne 2012 sera l'occasion d'éclaircir certains de ces points et de proposer une mise en perspective importante pour le TAL.

J'ai tenté systématiquement d'avoir recours dans ces quatre parties à des exemples des différents projets auxquels j'ai participé. La liste exhaustive (avec notamment les partenaires impliqués) est à la fin du document (page 235). La description détaillée des problématiques et des données en est faite au fil du texte, mais pas toujours de façon complète.

Première partie

Aperçu de mon parcours : une histoire personnelle dans des histoires collectives (informatique, TAL, linguistique)

Cette première partie est une évocation de mon parcours, au cours de laquelle je vais préciser les différents contextes et collaborations au sein desquels se sont déroulées mes activités de chercheur, tout en identifiant à chaque fois l'impact que cela a eu sur ma vision du domaine et le développement de mes centres d'intérêt. Je me contenterai donc dans la plupart des cas de survoler ces travaux parfois assez éloignés dans le temps, mais il paraît intéressant de voir les traces que ceux-ci ont laissées dans ma vision actuelle. De ce fait, je m'excuse platement auprès des nombreux collègues dont j'ai pu à l'occasion croiser la route, car tous ne se retrouveront pas cités explicitement dans ces quelques pages. Cela ne veut bien entendu pas dire que je les ai oubliés.

J'ai également tenté de replacer mes travaux et mes rencontres dans le cadre plus général de l'évolution du TAL et de la linguistique de corpus. Cette recherche d'un positionnement dans un paysage global m'a conduit à identifier quelques événements marquants de la vie de la communauté académique, mais aussi des phases de transition qui ont vu les intérêts scientifiques se concentrer sur de nouvelles méthodes ou de nouveaux objets au fil de leur apparition.

La frise de la figure 0.1 propose ainsi une visualisation, le long de l'axe temporel, de quelques points que j'estime importants par rapport aux propos que je vais tenir dans cette première partie, et développer dans les suivantes.

Puisque la trame du début de ce mémoire est essentiellement biographique, j'ai indiqué dans la partie basse de cette chronologie les principaux événements de ma carrière, en indiquant à la fois les étapes de ma formation et les domaines de ma recherche dans l'ordre où je les ai développés. Dans la partie haute j'ai indiqué quelques points plus généraux : la mise en place des institutions de la vie académique du TAL (associations, laboratoires, formations universitaires). Les trois « nuages » situés près de l'axe temporel correspondent à des tournants importants du monde de la recherche en TAL, et identifient des périodes de transition où l'on a vu en quelques années les pratiques et les centres d'intérêt se développer autour de préoccupations nouvelles : le premier tournant concerne le développement et la montée en puissance des méthodes quantitatives en TAL (systèmes de traitement probabilistes, méthodes par apprentissage) au détriment des modèles symboliques et des grammaires issues du mouvement générativiste. La seconde concerne à la fois le TAL et la linguistique dans son ensemble, avec le développement des grands corpus numériques qui sont devenu des objets incontournables à la fois pour l'étude du langage et pour le développement des applications informatisées. Le dernier concerne l'utilisation des ressources langagières nouvelles et de plus en plus massives mises à disposition des linguistes par le développement du Web. Chacune de ces phases de transition est évoquée plus en détails dans des chapitres de ce mémoire (respectivement les chapitres 8, 3 et 4).

On remarquera que mes propres travaux et centres d'intérêt n'ont pas toujours été synchrones avec ces évolutions : je ne me suis intéressé aux méthodes quantitatives que bien après leur apparition au premier rang de la scène des travaux en TAL, alors que je me suis

assez rapidement penché sur les différentes exploitations du Web comme source de données langagières.

Les choix des détails visuels de cette frise sont eux aussi arbitraires : par exemple la position verticale des différents éléments n'est pas significative. Je n'ai cherché dans ce positionnement que la lisibilité, même si au final la disposition en diagonale montante ouvre la porte à différentes interprétations.

Ma volonté de représenter ce point de vue global par un graphique trouvera un écho plus loin dans ce mémoire, lorsque j'aborderai explicitement les techniques de visualisation (chapitre 5), et je considère donc ce tout premier schéma comme un des éléments visuels d'une petite collection que j'ai pu construire dans des contextes au final assez variés.

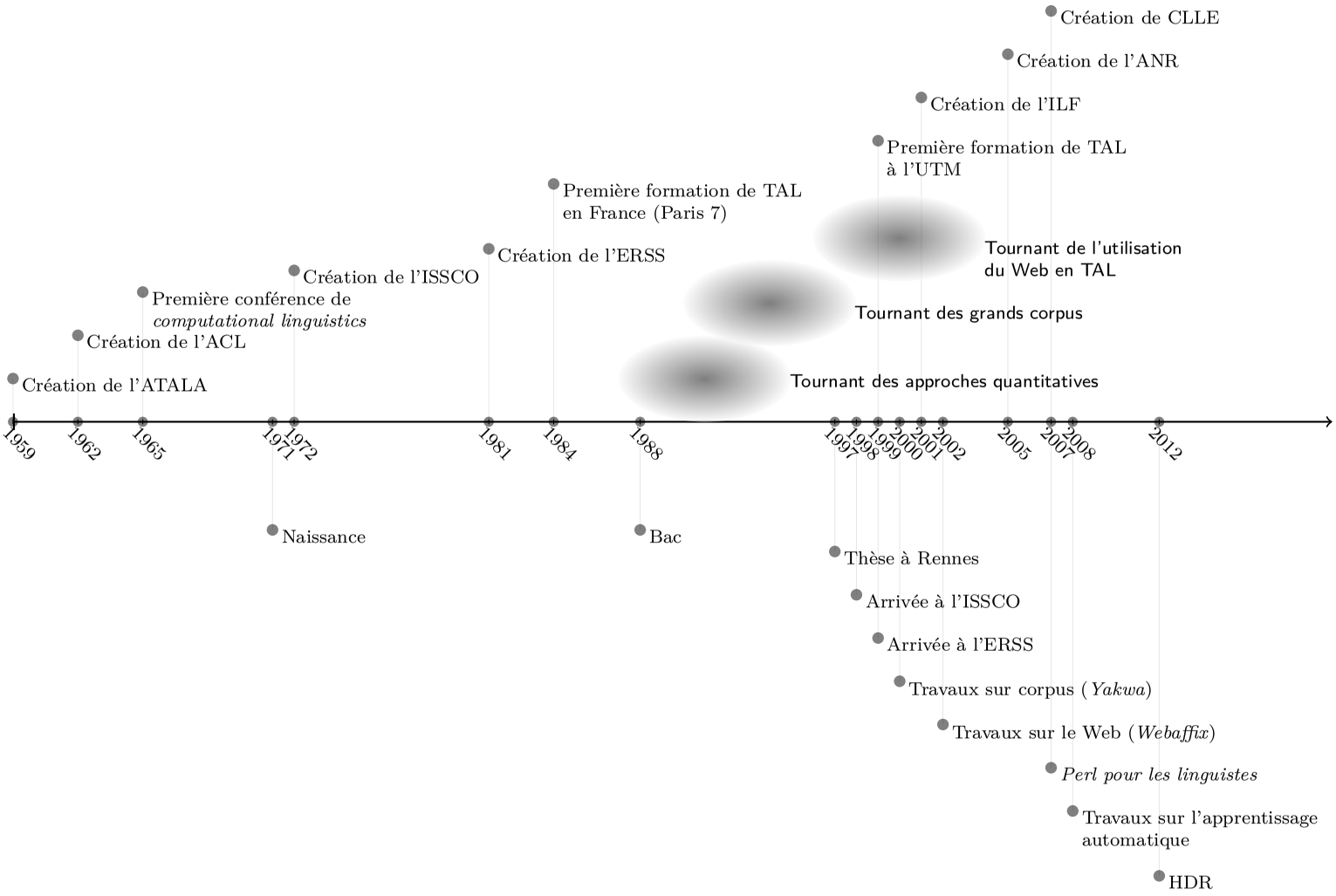

FIGURE 0.1 – Frise chronologique de quelques événements marquants

Chapitre 1

Doctorat et alentours : de l'informatique aux sciences du langage

Je commencerai ici par retracer les premières étapes de mon parcours de chercheur, en dégageant les principales thématiques de ma recherche initiale et les caractéristiques méthodologiques de mes premiers travaux. Cette période couvre la fin de mes études de second cycle (1993), l'obtention de mon doctorat en informatique (en 1997), mes deux années d'ATER à l'Université de Bretagne Occidentale (1997 et 1998) et mon activité de chercheur en linguistique informatique à l'ISSCO (1998 et 1999).

1.1 Premiers pas vers le langage (1993-98)

J'ai fait mes premiers pas dans le monde de la recherche en Bretagne, que ce soit à l'École Nationale Supérieure des Télécommunications de Bretagne (désormais nommée plus prosaïquement *Telecom Bretagne*), où j'ai effectué ma formation d'ingénieur et mon doctorat, à l'Université de Rennes 1 où j'ai effectué mon DEA, et enfin à l'Université de Bretagne Occidentale (UBO) où j'ai effectué des enseignements en tant que chargé de cours et ATER, et où j'ai rencontré mes premiers linguistes.

1.1.1 Éloignement de l'ingénierie en informatique

Ma formation initiale est celle d'un ingénieur en télécommunications. Ce type de formation n'est pas totalement sans rapport avec les choses langagières, recouvrait notamment le traitement du signal (dont la parole), la théorie de l'information, les modèles de codage/décodage, etc. Mais les points communs le plus importants entre ces enseignements et mon activité postérieure proviennent surtout de ma spécialisation en informatique (en plus de quelques réminiscences des modélisations stochastiques utilisées en traitement de signal) : théorie des langages, modélisation, bases de données, programmation.

Même après avoir choisi la voie de l'informatique, j'ai découvert assez rapidement un manque d'appétit pour les problématiques alors centrales dans ce domaine à l'époque (réseau, calculs distribués, génie logiciel, etc.), mais fort heureusement l'école venait juste de mettre en place une nouvelle filière baptisée à l'époque « Intelligence artificielle et sciences

cognitives », s'appuyant sur un laboratoire, le LIASC (Laboratoire d'Intelligence Artificielle et Sciences Cognitives). Cette formation et ce laboratoire avaient vu le jour grâce à l'arrivée de Jean-Pierre Barthélemy en provenance de Telecom Paris, où un tel cursus avait déjà été mis en place. C'est dans le cadre de cette formation (en parallèle avec un DEA d'informatique) que j'ai été confronté aux sciences cognitives, dans un joyeux panorama de toutes ces sciences humaines qui y étaient survolées (surtout la psychologie cognitive, mais également la linguistique, elle aussi toutefois parfumée de cognitivisme). L'époque était intéressante sur bien des points, et en constante évolution [1] : l'intelligence artificielle commençait sérieusement à perdre de sa superbe (certains voulaient d'ailleurs recycler le sigle IA comme signifiant Informatique Avancée), et les sciences cognitives n'étaient pas encore dominées par les neurosciences (de mon point de vue, c'étaient les linguistes que l'on trouvait alors en nombre dans les rassemblements étiquetés sciences cognitives). La part belle était donnée dans mon entourage à la représentation des connaissances, aux systèmes d'aide à la décision (la spécialité de J.-P. Barthélemy) et aux problématiques liées aux interfaces homme-machine. Un tournant important pour moi était celui qui prenait de la distance par rapport aux systèmes artificiels censés imiter le comportement cognitif humain (comme les fameux systèmes experts) pour aller vers des dispositifs qui assistent l'homme dans ses activités de haut niveau (Visetti, 1991). Le vocable naissant à l'époque était celui de *système anthropocentré*, c'est-à-dire des systèmes de représentation de connaissances, avec un mécanisme d'inférence limité et interactif, qui comme leur nom l'indique placent l'homme au cœur du système.

Les thématiques et terrains d'activités du LIASC étaient déjà très variés. Je citerai comme exemple : l'extraction de l'expertise d'un décideur (observation, simulation de situations), le développement d'interfaces homme-machine complexes pour assister des experts (visualisation en 3D pour la médecine), l'analyse de données en sciences sociales (études des déplacements pour un projet d'urbanisme), étude du comportement des utilisateurs (jeux vidéos), etc. Le tout passait généralement par une phase de construction d'un modèle mathématique spécifique (la plupart des membres du laboratoire étaient à la base des mathématiciens, J.-P. Barthélemy le premier), parfois très complexe mais gardant toujours un lien très ténu avec une implantation informatique (treillis, graphes, structures hiérarchiques).

Comme on le voit, le langage et la linguistique n'étaient guère représentés dans ces activités. C'était pourtant le domaine qui m'intéressait nettement le plus, et il se trouve que c'était également dans cette direction que voulait s'orienter Ioannis Kanellos, mon encadrant de thèse, alors fraîchement issu de la sémantique cognitive après un doctorat avec J.-P. Desclés. C'est lui qui avait découvert les travaux de François Rastier et qui m'a convaincu d'y consacrer mon travail de doctorat.

1.1.2 Entrée dans la linguistique : la sémantique interprétative de François Rastier

Avant de commencer à me plonger dans cette thèse (Tanguy, 1997), je n'avais à mon actif que quelques lectures de travaux de linguistique (mes premières lectures étaient Langacker, Chomsky, Desclés et quelques travaux anciens de psycholinguistique dont j'ai perdu les références). Ma véritable entrée dans ce champ se fit donc par la lecture exhaustive et intensive de la Sémantique Interprétative (ci-après SI), présentée dans (Rastier, 1987).

1. Il peut être intéressant de voir l'évolution terminologique au fil du temps : l'option et le département responsable se sont appelés CHMEST (Coopération Homme-Machine et Ergonomie des Systèmes de Télécommunications) puis actuellement LUSSI (Logique des Usages, Sciences Sociales et de l'Information).

Il va de soi que les premiers moments furent extrêmement difficiles : la masse de connaissances mobilisées qui caractérisent les travaux de Rastier est toujours aussi impressionnante bien des années plus tard, et c'est avec beaucoup de naïveté et d'humilité que je passais les premiers mois à tenter de combler les gouffres d'érudition qui me séparaient de cet univers. Je devais mon salut à peu de choses : la familiarité avec les modèles classiques de représentation des connaissances critiqués par Rastier, la clarté et l'intérêt des exemples issus d'œuvres littéraires que je connaissais pour certaines assez bien et enfin une certaine fascination pour le style de l'auteur.

Entrer dans la linguistique par la porte de la sémantique des textes a eu un ensemble de conséquences dans ma vision de la discipline. Pour commencer, c'est par cet angle que j'en découvrais les différentes branches et les niveaux d'analyse : les domaines confrontés plus directement au matériau brut (morphologie et syntaxe) m'apparurent à la fois fastidieux et très limités dans leur portée. Ce fut également le cas pour la sémantique lexicale, elle aussi apparaissant comme statique et rudimentaire dans ma lecture de Rastier. Cette condescendance facile m'entraîna rapidement à délaisser les traitements automatiques correspondants, que je ne découvris que quelques années plus tard en prenant véritablement pied dans le domaine du TAL.

À ce stade la construction d'un système informatique basé sur cette théorie me paraissait totalement inaccessible (c'était quand même le but de mon travail). Le programme rastiérien rejetait la possibilité du calcul du sens comme une émergence à partir de données génériques préétablies, et le paradigme classique de l'IA n'apparaissait simplement pas applicable.

En revanche, la modélisation (au sens de la représentation) elle-même était tout à fait envisageable, le gros du travail étant déjà fait par l'inscription de la SI dans le courant structuraliste, même s'il restait un grand nombre d'imprécisions qui allaient nécessiter des clarifications et des compléments.

De cette immersion dans une œuvre aussi marquante, je garde très clairement un ensemble de prises de positions futures sur les questions linguistiques, alors que j'étais à des années de me considérer comme un linguiste. Des principes rastiériens je retiendrai à ce stade les points suivants :

- La nécessité de travailler sur des textes et des données réelles et complètes. Si je n'ai pas eu l'occasion par la suite de travailler sérieusement sur des textes littéraires, j'ai toujours préféré utiliser des textes complets et pour lesquels on dispose du maximum d'informations sur leur origine.
- La primauté du contexte global sur les phénomènes locaux. Cet aphorisme me paraissait à l'époque plus destiné à se distinguer des travaux de l'IA sur des données artificielles et décontextualisées (les phrases-jouets manipulées par certains outils, comme les fameuses *donkey-sentences* chères à la formalisation logique). Dans mes différents travaux, et même en me confrontant à d'autres domaines que la sémantique, je ne pense pas par la suite avoir dérogé à ce principe.
- Corollaire du point précédent, l'impact du genre textuel sur toute considération linguistique particulière, et par là l'impossibilité d'envisager un traitement unique pour tout type de données.
- La multiplicité inévitable des interprétations et des significations de tout passage textuel et la conviction qu'un sens unique et calculable est totalement illusoire.

1.1.3 Propositions : l'outil informatique pour assister l'analyse des textes

Les prises de position de la théorie rastiérienne et le contexte scientifique dans lequel se déroulait mon doctorat me firent donc privilégier la représentation d'un sens déjà construit, et l'explicitation de celui-ci. Puisque la somme de connaissances nécessaires pour « faire tourner » la sémantique interprétative sur un texte est immense, et que la subjectivité du parcours interprétatif y est revendiquée, ma proposition fut donc, à travers un modèle très simple et directement calqué des propositions de Rastier, de faire jouer à l'outil informatique le rôle d'un *assistant à l'explicitation d'un parcours interprétatif*. La seule et unique source de connaissance dans ce système était l'utilisateur et sa compréhension du texte et du contexte (et non pas par exemple des bases de traits sémantiques prédéfinies). Les seules fonctions propres à l'outil informatique étaient donc le stockage et la représentation d'une annotation sémantique, la vérification de sa cohérence et, de façon peut-être moins convaincante, la recherche de complétude du système par sollicitation interactive de l'utilisateur. Autrement dit, mon approche se concentrait plus sur le rôle de l'outil informatique dans le processus interprétatif lui-même, que sur les résultats. En ce sens, mon travail se démarquait à la fois des travaux de la linguistique computationnelle qui visait à simuler ce processus, mais aussi de l'outillage classique des analyses de textes (concordanciers, approches lexicométriques) qui ne proposait qu'un accès facilité au matériau textuel et restait extérieur à l'activité sémantique elle-même.

Le principal retour pour la théorie elle-même fut la spécification de certains aspects purement formels (la nécessité de la modélisation m'a conduit à préciser certains éléments du formalisme de la sémantique interprétative, et même à inventer quelques concepts intermédiaires (comme le peu élégamment nommé *spécème*[2] apparemment repris par d'autres selon Google).

Si je devais résumer le fonctionnement du programme (baptisé PASTEL, pour « Programme d'aide à l'Analyse Sémantique des TExtes, même Littéraires ») qui constitue le point final de cette thèse, je le ferais ainsi :
Étant donné un texte, qu'on choisira court et riche en figures (par exemple un poème ou un article du *Canard Enchaîné*, pour reprendre les deux seuls exemples testés) et un utilisateur bien disposé, l'outil propose à ce dernier une interface graphique lui permettant de définir des classes sémantiques (ne rentrons pas ici dans les nuances de la SI) et d'y attribuer des mots ou expressions relevées dans le texte, occurrence par occurrence. Une fois ces premières classes établies, l'utilisateur toujours motivé se voyait demander par un processus itératif de qualifier ce qui différenciait les éléments de ces classes en attribuant des traits sémantiques (sèmes spécifiques) à chacun, jusqu'à obtenir une caractérisation sémique unique de chaque élément signifiant. De nouvelles sous-classes pouvaient ainsi émerger, à l'intérieur des premières ou transversalement à celles-ci, et ré-enclencher le processus itératif. Le système passait donc à travers de telles phases de dialogue d'un état stable d'une représentation sémantique à une autre, plus complète, jusqu'à épuisement de la bonne disposition de l'utilisateur.

Un autre retour proposé par la machine, et sur lequel je reviendrai au chapitre 5 est la représentation graphique, possible en fin d'analyse, de la disposition des classes sémantiques dans le texte (isotopies). La figure 1.1 montre une telle représentation (avec l'outil d'époque) : les deux lignes de symboles baptisées *Forme 1* et *Forme 2* indiquent par des étoiles les lieux relatifs du texte où les deux classes sémantiques indiquées au-dessus s'instancient. On peut en

2. Ce concept pointu correspond au support formel d'un sème spécifique, i.e. la relation d'opposition qui fonde la différence entre les sens de deux lexèmes proches.

déduire par exemple des dispositions relatives (juxtaposition, superposition, alternance, etc.)

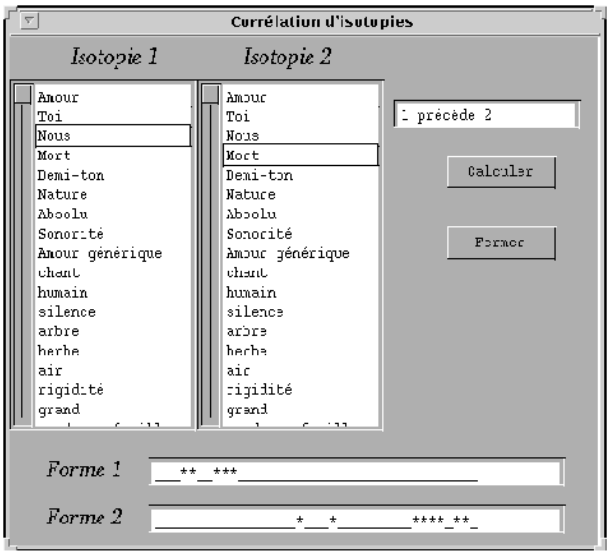

FIGURE 1.1 – Disposition d'isotopies

Cette tentative était surtout motivée par un besoin de fonctionnalités supplémentaires proposant un retour vers l'utilisateur (comme pour récompenser les efforts que celui-ci a dû fournir pour satisfaire le système). Mais elle ouvrait également la voie vers des notions plus complexes de la sémantique interprétative, notamment la composante tactique introduite dans Rastier (1989).

1.1.4 Modèles informatiques pour la linguistique

Si le modèle que j'ai proposé était d'une simplicité extrême et ne s'inspirait que de la théorie des ensembles, les travaux que j'ai en petite partie partagés avec Théodore Thlivitis pour sa thèse (Thlivitis, 1998) sur la modélisation de la dimension intertextuelle dans la sémantique interprétative faisaient, eux, appel à la technologie UML [3]. Ces travaux sont notables car ils montrent l'enrichissement que peut fournir une connaissance des techniques issues de la science informatique plus *hard core* (ici les techniques de conception de logiciel) dans l'explicitation d'un modèle linguistique. Thlivitis a ainsi pu développer des concepts formels assez poussés pour aborder le monde de l'intertexte comme les lectures multiples et les influences des interprétations d'un texte sur un autre, dans un modèle bien plus complexe. Malheureusement, un tel modèle nécessitait une masse de connaissances à représenter encore plus importante pour pouvoir démontrer son application.

En parallèle de mon travail de thèse, j'ai eu l'occasion d'utiliser mes compétences de modélisation lors d'un travail collaboratif très enrichissant avec Michel Schmouchkovitch, psychiatre qui suivait à l'époque le séminaire de linguistique organisé à Brest par Michèle Noailly. Dans son travail de praticien hospitalier, il avait été confronté à une patiente qui lui avait rapporté avoir construit un système (ou plutôt un code) attribuant des significations

3. Pour *Unified Modelling Language*, un langage de représentation des modèles informatiques utilisé pour le développement des programmes orientés objet.

très spécifiques à chaque lettre et chiffre, ce qui lui permettait d'interpréter des noms, plaques minéralogiques, numéros, etc. Mon travail a consisté à construire, à partir des consultations enregistrées, le graphe de ce code, afin notamment d'y repérer les classes de significations, les types de motivations et les (rares) incohérences du système. Ce travail qui m'a paru très simple sur le plan technique a été extrêmement bien reçu et déclaré très éclairant par le praticien, en lui apportant une vue synthétique (sous la forme d'un graphe) donnant une vue d'ensemble sur la complexité du système, mais sans impact notable sur la santé de la patiente (Schmouchkovitch *et al.*, 1998). Je présente plus en détails cet épisode en section 5.1.1.1.

1.1.5 Suite et fin

Ma thèse fut, à la soutenance, qualifiée d'« orthogonale », n'entrant pas dans les paradigmes classiques de l'interaction entre la linguistique et l'informatique (ce qui me remplit de fierté avant de découvrir que c'était aussi un handicap pour mon insertion). Au final, je n'étais toutefois pas seul, plusieurs autres informaticiens gravitaient autour des travaux de Rastier : Marc Cavazza (Rastier *et al.*, 1994), Pierre Beust (Beust, 1998), Bénédicte Pincemin (Bommier-Pincemin, 1999), etc. Certains d'entre eux, et d'autres à leur suite ont repris le flambeau et poussé bien plus loin l'utilisation des idées de Rastier dans le traitement automatique du langage. C'est sans doute Mathieu Valette (Valette, 2009) qui à ma connaissance est allé le plus loin, et a proposé les systèmes les plus complets et opérationnalisés reprenant les principes de l'analyse sémique et la projection/induction de traits sémantiques dans les textes.

Il était par contre bien clair que l'outil que j'avais développé n'avait au plus qu'une valeur de preuve de l'opérationnalité du modèle de la SI que j'avais proposé. Malgré mes tentatives pour essayer de le faire utiliser par d'autres, notamment des étudiants de littérature, il était très éloignés des besoins réels de ceux-ci, j'y reviendrai par la suite.

Je pense qu'à la fin de ce travail de doctorat, j'étais dans une position idéologique assez extrême, et rejetais assez fortement toute possibilité d'une projection de connaissances lexicales génériques sur un texte donné, en suivant le principe que l'interprétation était libre et non contrainte, et qu'un parcours interprétatif individuel sur un texte précis était à même d'effacer (ou du moins de virtualiser) toute connotation d'une lexie, ce qui était suffisant pour démontrer l'inutilité de classes préconstruites. F. Rastier avait dit d'ailleurs lors de ma soutenance que j'étais (je cite de mémoire) « plus rastiérien que lui, mais que je ne devais pas m'inquiéter, cela passe avec l'âge ». Il avait, comme d'habitude, totalement raison.

1.2 Outillage informatique de la géolinguistique bretonne : mérites de la visualisation et dangers de l'automatisation

De façon totalement décorrélée de mes travaux de doctorat, mon entrée timide dans le monde de la linguistique extra-rastiérienne se fit à cette époque par le biais d'une collaboration avec Jean Le Dû, spécialiste de géolinguistique bretonne au Centre de Recherche Bretonne et Celtique de l'université de Brest. Nous nous étions rencontrés là aussi lors d'un des séminaires organisés par Michèle Noailly, au cours duquel il présenta avec son collègue Yves Le Berre une discipline dont je n'avais jamais entendu parler, la géolinguistique. Comme d'autres dialectologues que je devais rencontrer plus tard à Toulouse (Jean-Louis Fossat, Guylaine Brun-Trigaud), leur travail se basait sur l'exploitation minutieuse de données d'enquêtes

recueillies plusieurs dizaines d'années auparavant à travers toute la Bretagne, et qui visaient l'inventaire des variations dialectales du breton (Le Dû, 2001). Disposant de ces données complexes (on dirait aujourd'hui *géolocalisées*) et somme toute assez volumineuses, leur outil principal d'investigation était les *isoglosses*, ou projections sur une carte géographique des différentes valeurs d'un trait linguistique (généralement les variations morphologiques d'une unité sémantique, ou les variations phonologiques d'une même unité lexicale) pour observer les zones ainsi délimitées sur le territoire. Ce travail, dont l'ampleur et la précision m'impressionnaient, passait par le repérage manuel à travers des listes de mots et de transcriptions phonétiques, et le dessin au crayon des zones correspondantes, avant d'arriver à l'interprétation, et notamment à la corrélation de ces zones linguistiques avec d'autres dimensions géographiques (frontières naturelles ou politiques, voies de communication, etc.). Ma culture d'ingénieur me poussa immédiatement à leur proposer l'assistance de l'outil informatique pour l'automatisation des tâches les plus mécaniques, et aboutit après quelques mois, et avec l'aide de collègues du laboratoire LIASC, à un prototype très grossier qui remplissait au final les fonctions voulues : pour un trait donné, exprimé par des expressions régulières projetées sur la base de données, un sous-ensemble de formes retranscrites étaient sélectionnées, leurs coordonnées spatiales identifiées, et les zones délimitées étaient alors dessinées automatiquement sur un fond de carte par des lignes colorées. Autrement dit, en quelques secondes un programme informatique réalisait (sans se tromper) un travail qui aurait pris plusieurs heures. L'enthousiasme de Jean Le Dû face à cet outil fut une des expériences les plus valorisantes de mes premières années de chercheur (j'allais heureusement en connaître quelques autres du même type par la suite comme on le verra). La figure 1.2 présente une telle carte, dans la toute première version du prototype. Y apparaissent les différentes réalisations pour exprimer *le bourg* en breton. Le rendu fut par la suite nettement amélioré grâce au travail de plusieurs étudiants, comme présenté dans (Kanellos *et al.*, 1999).

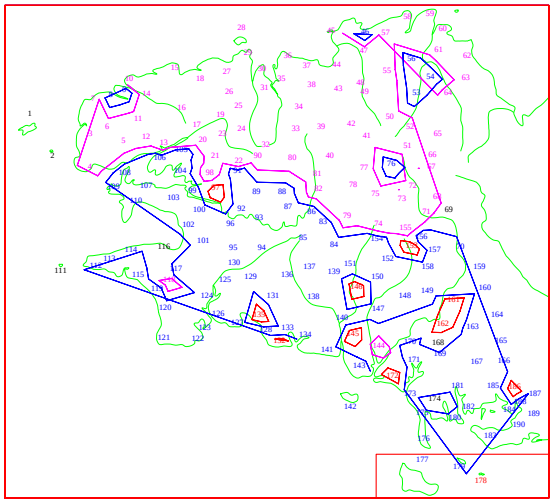

FIGURE 1.2 – Carte géolinguistique de la Bretagne (formes *le bourg*), générée automatiquement
Les lignes colorées (*isoglosses*) délimitent les zones géographiques pour lesquelles les formes prononcées sont similaires

J'ai réalisé à cette occasion que des compétences techniques relativement simples pouvaient s'appliquer directement à des travaux dans des disciplines assez pointues, pour peu que l'on prenne la peine d'en comprendre les tenants et les aboutissants. Un autre point intéressant à noter était les doutes que j'émettais alors sur les biais apportés par ces mêmes méthodes automatisées qui étaient recherchées dans ce type de collaboration : j'étais persuadé de l'intérêt pour le chercheur de faire ce travail manuel sur ses données, seul à même de déclencher chez lui l'intuition qui le fera chercher au bon endroit. Tout outil de recherche automatique (même comme ici une petite expression régulière projetée sur des chaînes isolées) a en effet tendance à éloigner l'utilisateur des données de départ. Dans la même veine de scepticisme, j'alertais également les utilisateurs sur le danger des choix et approximations inhérents aux algorithmes utilisés. Certaines découvertes de la géolinguistique concernaient en effet les points aberrants qui ressortaient d'un tracé donné : si certains n'étaient que des erreurs dues à l'enquête ou à la transcription, d'autres étaient par contre des indices isolés d'un état de la langue qui ne transparaissait que par eux (par exemple des localités périphériques ayant connu moins rapidement que d'autres les évolutions du breton). La méthode automatique, elle, devait par contre prendre une décision dans le tracé, et risquait fortement de faire purement et simplement disparaître de la synthèse des informations vitales. Je reviendrai dans la suite sur ces deux doutes formulés à l'époque et applicables à d'autres situations d'intervention de l'outil informatique.

Ce travail a également été important pour moi, bien que je ne m'en sois rendu compte que bien plus tard, en ce sens qu'il me faisait entrer dans la problématique de la visualisation des données. Présenter les résultats d'une analyse sur un objet visuel, surtout aussi commun qu'une carte géographique (et donc facilement interprétable, les codes de lecture faisant partie d'une culture universelle) est une excellente façon de partager des résultats et des questions. J'ai pu vérifier par la suite qu'une représentation visuelle est un excellent outil de collaboration interdisciplinaire.

1.3 Séjour à l'ISSCO : découverte du TAL (1998-99)

L'étape suivante de mon parcours m'a conduit à quitter la Bretagne et la France pour occuper à Genève un poste de chercheur en linguistique informatique à l'ISSCO d'avril 1998 à septembre 1999.

1.3.1 Un « vrai » laboratoire de TAL

L'ISSCO (*Istituto per gli Studi Semantici e COgnitivi*) est un institut de recherche qui avait depuis sa création un statut de fondation (financée par feu le richissime philanthrope Dalle Molle) et qui était alors dans une phase de transition qui allait aboutir à son rattachement à l'université de Genève, et plus précisément à l'École de Traduction et d'Interprétation (ETI), où il recevrait son nom actuel de TIM (Traitement Informatique Multilingue).

Malgré ces changements de nom et de statut, TIM/ISSCO n'en reste pas moins un des plus anciens laboratoires de TAL d'Europe (fondé en 1972, il a à peu près mon âge), et a vu défiler entre ses murs de très nombreux chercheurs de toutes les nationalités, dont plusieurs monstres sacrés. Son fonctionnement à l'époque, sous la direction de Maghi King, était entièrement orienté vers des projets de recherche financés par la communauté européenne et le fonds national suisse de la recherche. Mon recrutement était lié à des besoins concernant principalement deux projets européens qui abordaient chacun un des deux grands domaines

d'excellence du laboratoire à l'époque : l'évaluation des outils et ressources du TAL d'une part, et le traitement multilingue d'autre part.

En ce qui concerne mon parcours personnel, il est important de noter que mon arrivée à l'ISSCO est mon premier contact véritable avec le monde du traitement automatique des langues : mes travaux précédents ne pouvaient alors pas véritablement être inscrits dans ce domaine, et je n'avais jusque là pas été confronté à des données massives génériques (lexiques, grammaires) ni à des outils d'annotation automatique (étiqueteurs, parseurs, etc.). Toutefois, mes compétences en programmation, ma connaissance théorique de la linguistique informatique, ainsi que les liens pouvant être établis entre mes travaux et la problématique de l'étiquetage sémantique (alors une des problématiques de certains membres de l'ISSCO comme Pierrette Bouillon) avaient permis à ma candidature d'être retenue.

1.3.2 Projet DiET : évaluation du TAL

Le premier projet sur lequel j'ai travaillé pendant mon séjour à l'ISSCO était le projet DiET (*Diagnosis and Evaluation Tools for Natural Language Processing*) dont l'objectif était le développement d'un banc de test pour un ensemble assez large d'outils de TAL. Ce projet était le prolongement d'un autre projet européen auquel avait participé l'ISSCO, TSNLP (*Test Suites for Natural Language Processing*, décrit dans (Lehmann *et al.*, 1996)), qui avait consisté en l'établissement de jeux de test (lexiques et phrases annotés) permettant de mesurer l'efficacité de différentes applications de TAL. La nouveauté du projet DiET résidait dans l'utilisation de techniques semi-automatisées de sélection, d'adaptation et d'extension de ces données de test (Bouillon *et al.*, 1999; Tanguy *et al.*, 1999a).

Mon rôle était de développer un système de profilage de textes pour aider à la sélection de jeux de test. Le postulat de départ était qu'une application donnée (un traducteur automatique, un correcteur grammatical, un vérificateur de langage contrôlé, etc.) est généralement conçue, notamment dans un milieu industriel, pour traiter un type de document assez précis, et pas nécessairement des textes tout venant. Ainsi, sur la base d'un jeu de test générique, certains items peuvent être pertinents (puisqu'ils correspondent à des phénomènes ou des configurations linguistiques qui seront rencontrés lors de l'utilisation de l'application) et d'autres pas. Le rôle du profilage de texte dans DiET était donc, pour un jeu de test et un corpus de textes représentatifs donnés, de sélectionner les seuls items du jeu qui sont présents dans le corpus.

Le cœur de ce module était constitué d'un moteur de recherche sur des textes, et d'un jeu de patrons correspondant aux phénomènes abordés dans les jeux de test. Le tout devait de plus être conçu pour être configuré par un utilisateur, en le laissant ajouter lui-même des items de test et des patrons correspondants. La cible centrale était toutefois constituée de jeux de test correspondant à la correction grammaticale, ainsi qu'à la résolution d'anaphore : les principaux phénomènes à repérer concernaient donc des configurations morpho-syntaxiques particulières (phrases interrogatives, négations, divers types de marqueurs anaphoriques, temps verbaux, etc.).

Pour répondre à cette demande, j'ai alors développé un système générique permettant de rechercher de telles configurations dans un texte préalablement étiqueté morphosyntaxiquement. Il était basé sur des automates à états finis, et sur un langage adhoc permettant leur définition par des expressions. Un exemple d'un tel patron est par exemple la séquence :

Déterminant + Nom + Adjectif + Conjonction de coordination + Adjectif

qui permet de trouver des configurations dans lesquelles se pose la question de l'accord d'adjectifs coordonnés pour lequel plusieurs phrases annotées existent dans les jeux de test prédéfinis. Les fonctionnalités de ce langage de patrons, présenté plus en détails dans (Rebeyrolle et Tanguy, 2000) et en section 3.2, comprenaient des contraintes sur la forme ou le lemme des unités lexicales (par exemple, finissant par *-ant*), des opérateurs de négation (par exemple, PAS un adjectif) et des fermetures (pour définir des séquences de taille quelconque de mots respectant une contrainte). En plus du langage lui-même, j'ai développé une interface graphique permettant à un utilisateur de définir ses propres patrons sans avoir à apprendre la syntaxe du langage formel ou le jeu d'étiquettes morphosyntaxiques.

Une autre fonctionnalité de l'outil de profilage était le calcul d'un ensemble de caractéristiques quantitatives directement extraites de l'analyse du texte, et qui devaient permettre d'identifier les principales caractéristiques de surface du texte (taille des phrases et des mots, répartition des parties du discours, richesse lexicale, etc.).

Ce module de profilage (alors appelé TPS pour Text Profiling System) fut ensuite intégré dans un immense programme Java dont les multiples fonctionnalités et l'architecture complexe (en réseau, autour d'un système de base de données) articulaient les autres composants du projet.

Le bilan du projet fut mitigé : si sur le plan technique celui-ci fut un succès qui fit la fierté de l'équipe internationale au sein de laquelle j'avais travaillé, sur le plan plus pratique de l'utilisabilité le plus simple est d'utiliser le terme consacré d'*usine à gaz*. J'appris par la suite que ce destin est malheureusement celui de nombreux projets de recherche de ce type, dans lesquels les besoins des utilisateurs finaux sont très rapidement enfouis sous des considérations techniques. De la bouche d'un des utilisateurs finaux (que j'ai rencontré des années plus tard), DiET avait même atteint sur ce plan un niveau de légende locale...

Sur le plan personnel, ce projet m'a appris énormément. En ingénierie tout d'abord : j'y ai appris la programmation en Java, diverses techniques informatiques (multi-programmation, mise en réseau, programmation événementielle, etc.) et le développement en équipe et à distance. Pour ce qui concerne le TAL, j'ai été confronté à l'étiquetage morphosyntaxique, à la gestion de données volumineuses, à des corpus et des lexiques dans plusieurs langues. Après DiET, je savais avec beaucoup plus de certitude ce qu'était le côté applicatif TAL, et les compétences que celui-ci exige. J'avais aussi appris à travailler en équipe et à gérer les conflits et tensions qui ne manquent pas d'apparaître.

Comme on le verra par la suite, l'outil que j'ai développé à cette occasion allait connaître une deuxième vie : si son utilisation dans le cadre du projet est restée marginale, il allait être mon point d'entrée pour un ensemble de travaux qui nécessitent d'explorer des textes pour y rechercher des configurations complexes, comme je le présente dans le chapitre suivant.

1.3.3 Projet IDOL : ingénierie multilingue

Le second projet auquel j'avais été affecté s'inscrivait dans le domaine du TAL multilingue, l'autre domaine d'excellence de l'ISSCO. Il ne s'agissait toutefois pas de traduction automatique mais assistée : je retrouvais dans ce cadre les positions théoriques abordées dans le cadre de ma thèse, puisqu'il s'agissait ici de développer une plate-forme d'aide à la traduction. Un tel système se limite au repérage de termes présents dans une base terminologique, à l'identification de segments de textes déjà traduits dans un autre document (mémoire de traduction), à la gestion de lexiques multilingues, le tout dans une interface graphique intégrée

à un traitement de texte. Si ce type d'applications était déjà bien développé à l'époque, et avait même atteint le niveau commercial, la plate-forme visée par le projet IDOL (IRS-based DOcument Localisation) avait la spécificité de s'attaquer à une langue alors très peu abordée par le TAL : l'arabe.

Mon rôle dans le projet était de développer un module de vérification de traduction, nommé TRACER (*TRAnslation CheckER*) qui, intégré à un logiciel de traitement de textes, devait proposer au traducteur un ensemble de mesures de vérification pour l'aider à identifier et corriger les erreurs de son travail.

Le module développé se basait dans un premier temps sur des techniques classiques dans le domaine de la traduction assistée : un programme d'alignement statistique de phrases basé sur l'algorithme proposé par Gale et Church (1993) proposait de façon automatique d'apparier les phrases du texte cible avec celles du texte source. Sur la base de cet alignement (et du déroulement de l'algorithme), un ensemble de vérifications simples pouvaient être faites, pour notamment vérifier si les tailles du texte et de ses sous-parties indiquaient un oubli ou non de passage à traduire. Dans un second temps, des ressources lexico-terminologiques translangues y étaient projetées, permettant là encore de vérifier dans quelle mesure la traduction respectait un référentiel.

Sur la base de cette seconde vérification, et au vu des faibles résultats obtenus par la recherche de correspondances normées entre termes traduits, j'ai proposé une approche qui s'inspirait directement de mes travaux de doctorat. Si l'opération de traduction, même technique, implique une certaine souplesse dans le choix des termes, elle doit néanmoins respecter les grandes structures sémantiques du texte à traduire. Et si, au lieu de rechercher des correspondances bi-univoques entre termes, on recherche des correspondances de classes sémantiques, on peut vérifier la traduction à un niveau plus global. Je proposai donc, sur la base de classes lexicales bilingues établies manuellement à partir d'un lexique de transfert, la comparaison des configurations textuelles de ces classes dans le texte source et dans sa traduction, en reprenant le principe greimasso-rastiérien des isotopies (Tanguy *et al.*, 1999b). Plus simplement, si l'on repère dans une ou plusieurs zones du texte source l'occurrence de plusieurs unités lexicales appartenant à une classe sémantique donnée, le même schéma doit plus ou moins être observable dans le texte cible. Plutôt que de comparer des points dans le bitexte, on recherche donc des schémas d'isotopies similaires.

Cette proposition reposait sur la représentation graphique de ces isotopies, en comparant visuellement les différents schémas isotopiques des deux textes, comme je le précise en section 5.2.1 (page 122).

Les principaux défauts de cette approche étaient sa limitation à un type d'erreur global visible à l'échelle du texte (essentiellement des passages non traduits, ou à la rigueur des inversions de passages), son manque de robustesse face à des erreurs cumulées, et la nécessité de disposer de ressources lexicales riches (construites manuellement pour les cas d'étude présentés dans le projet).

Là encore, le bilan de ce projet fut mitigé, mais les résultats furent moins disproportionnés par rapport aux efforts fournis. Il constitua pour moi le premier contact avec l'ingénierie multilingue, dans sa version d'assistance à l'utilisateur, ce qui réconfortait mes convictions initiales sur le rôle de l'outil informatique face au langage. Par contre, il marqua également pour moi le début d'un éloignement avec les positions (parfois extrémistes) que j'avais prises dans le cadre de ma thèse en me confrontant à la nécessité d'intégrer des ressources linguistiques externes pour répondre à un besoin réel.

En plus de ces travaux dans le cadre de projets de grande envergure, mes activités à l'ISSCO ont abordé ponctuellement d'autres aspects du TAL : l'étiquetage morphosyntaxique (son paramétrage et son évaluation), les lexiques morphologiques, les dictionnaires informatisés, la désambiguïsation sémantique, etc. En plus de cette diversité d'activité, un des atouts indéniables de l'ISSCO était sa position internationale, à la fois par ses collaborations que par le cosmopolitisme de ses membres. Tout cela a fait que ce passage somme toute assez court a énormément élargi mon horizon de chercheur.

Chapitre 2

Arrivée à l'ERSS : de l'usage du TAL dans un laboratoire de linguistique

L'étape suivante (et la dernière à ce jour) de mon parcours académique concerne la position que j'occupe actuellement, d'enseignant-chercheur en sciences du langage à Toulouse, en tant que membre de l'ERSS (désormais CLLE). J'y suis arrivé en 1999, pour un retour en France que j'avais initialement envisagé plus tardif (j'avais pris goût aux projets européens, et envisageais initialement de devenir une sorte de mercenaire-consultant des projets en TAL, ayant eu plusieurs propositions en ce sens de la part des partenaires avec lesquels j'avais collaboré). D'un autre côté, le contexte du poste auquel je postulai était de ceux qui ne se refusent pas, et l'activité d'enseignement commençait à me manquer cruellement.

Ce chapitre précise l'évolution de mes travaux dans cette équipe de recherche, qui constituent l'essentiel de ce que je détaillerai dans ce mémoire. J'insisterai donc plus ici sur les motivations et l'insertion dans mon univers actuel.

2.1 Immersion dans la linguistique

J'ai donc atterri à la rentrée 1999 dans un monde alors très nouveau pour moi, pour plusieurs raisons. Premièrement, il s'agissait d'un grand laboratoire (par rapport à ceux que j'avais connus), deuxièmement, c'était un laboratoire du CNRS, structure que je ne connaissais absolument pas et enfin, il s'agissait d'un laboratoire de linguistique « pure ». Heureusement, plusieurs de mes collègues avaient des parcours très similaires au mien : Cécile Fabre, avec qui j'ai le plaisir de partager depuis toutes ces années mes activités d'enseignement en plus de ma recherche, Didier Bourigault, qui était arrivé à l'ERSS en même temps que moi, et que j'avais déjà croisé dans la sphère rastiérienne, et Nabil Hathout, dont je faisais alors la connaissance et avec qui je partage le bureau et une grande part de mes activités de recherche depuis le début, bien que ses thématiques de recherche me fussent alors totalement inconnues (la morphologie, les dictionnaires). Mes premières activités de recherche ont eu lieu dans le cadre de l'opération que dirigeait alors Anne Condamines, baptisée « Sémantique et corpus ». Pendant les premières années, j'ai ainsi pu me familiariser avec d'autres approches de la sémantique que celles que je connaissais, et découvrir la linguistique de corpus. Je dois notamment cet apprentissage aux dialogues avec Anne Condamines, Marie-Paule Péry-

Woodley et Cécile Fabre qui ont très rapidement su faire le lien entre mon expérience et les usages des outils informatiques pour l'investigation des corpus.

Un point très important à noter à ce stade est que je débarquais avec une culture nouvellement acquise de recherche sur projet, et après avoir laissé derrière moi mes travaux de thèse depuis deux ans. Autrement dit, sans projet de recherche spécifique, mais bien décidé à collaborer avec ces nouveaux collègues dont j'avais à apprendre les objets d'études, les théories, les méthodes de travail et bien entendu les questionnements techniques, méthodologiques ou scientifiques auxquels j'aurais pu répondre.

2.2 Outillage de la linguistique

Une grande partie de mon activité à l'ERSS a été consacrée à la collaboration avec différents membres autour de la problématique générale de l'exploitation informatique des données langagières. Je me contenterai ici de délimiter les principaux axes de ce travail, dont les détails et les spécificités seront présentés dans la suite de ce mémoire.

2.2.1 Fouiller les corpus

La place des corpus à l'ERSS est absolument centrale depuis le début, et ce pour l'ensemble des niveaux de description. Le laboratoire est d'ailleurs dépositaire d'une grande collection de corpus électroniques, très variés (mais malheureusement pas tous diffusables), et je n'ai vu que très peu de travaux réalisés par ses membres qui ne s'appuient pas sur l'un d'entre eux. Cette culture ambiante fait que les usages des corpus sont eux aussi très variés, comme on le verra en partie dans ce mémoire pour les travaux auxquels j'ai participé. Les membres de l'opération *Sémantique et corpus* ont de plus mené un ensemble de réflexions sur l'usage des corpus en linguistique (Condamines *et al.*, 1999; Péry-Woodley, 1995; Condamines, 2005b).

Pour ce qui est de mes travaux, le premier axe concerne l'exploitation de corpus électroniques, principalement pour y rechercher des structures de surface particulières, définies par le biais de patrons morphosyntaxiques. L'exemple le plus complet de ce travail est ma collaboration avec Josette Rebeyrolle, qui terminait alors son doctorat (Rebeyrolle, 2000) sous la direction d'Andrée Borillo et Anne Condamines sur les énoncés définitoires (du type *On appelle X le Y qui Z* ou *Le terme X désigne le Y qui Z*) ; je détaille ces aspects en section 3.1.2.2 (page 58). Mon rôle dans l'investigation linguistique fut assez limité, vu que j'arrivais après la phase principale qui avait conduit Josette Rebeyrolle à définir formellement les séries de patrons de surface qui délimitaient les différents types de définition de termes trouvés dans son corpus d'étude. Elle avait principalement utilisé pour l'exploration de ses données le logiciel SATO (Daoust, 1996), alors un des rares outils capables de manipuler simplement des patrons morphosyntaxiques, mais limité par l'absence de désambiguïsation catégorielle des unités lexicales. Les besoins qu'elle exprimait alors concernaient principalement une définition plus complète de ces patrons, et la mesure de leur efficacité à repérer les énoncés ciblés. Parallèlement, les travaux d'Anne Condamines rejoignaient les mêmes préoccupations, et concernaient différents types de marqueurs relationnels qu'elle définissait pour l'extraction de liens sémantiques entre termes. Ces différentes préoccupations convergeaient donc vers le besoin d'un outil capable de projeter des patrons morphosyntaxiques sur des textes étiquetés, s'appuyant sur des informations simples, mais en les articulant parfois de façon complexe. Il me suffit alors de donner une seconde vie au module de profilage du projet DiET dont les fonctionnalités correspondaient exactement à ces besoins (voir section 1.3.2, page 35). Pour je

ne sais plus quelle raison, l'outil fut nommé Yakwa. Son interface graphique fut complexifiée, et longuement testée par les collègues qu'elle visait comme utilisateurs, et certaines fonctionnalités de recherche très spécifiques furent ajoutées (mémorisation, recherche de répétitions, prise en compte de critères de position dans le texte). Il fut utilisé à l'ERSS pour un certain nombre de travaux que je n'énumérerai pas ici, et même utilisé à l'extérieur malgré ses exigences techniques sur le plan du système informatique et le manque d'investissement de ma part pour le rendre facilement installable. Cette période fut pour moi l'occasion de collaborer avec un grand nombre de chercheurs et doctorants, à Toulouse comme ailleurs, et de me familiariser avec les usages en linguistique de corpus et en linguistique appliquée. Ma participation apportait à ces travaux un outillage spécifique, mais également une méthodologie dans la définition des modes de recherche.

Cette activité a bien entendu évolué, avec notamment la plus grande disponibilité d'analyseurs syntaxiques qui augmentent les capacités de recherche mais rendent la tâche plus complexe techniquement et humainement, la montée en volume des données avec la mise à disposition de grands corpus (archives de journaux, base Frantext), et le développement d'outils plus performants que les premiers prototypes. Toutefois, la principale évolution est sans doute le fait que ces configurations particulières recherchées dans des corpus sont maintenant beaucoup plus utilisées comme briques de base pour des analyses plus complexes que comme produit final (si l'on met de côté l'analyse linguistique proprement dite). Autrement dit, c'est plus le TAL lui-même qui est le client de ces systèmes de recherche de patrons morpho-syntaxiques, et ils deviennent dans certains cas la porte d'entrée qui permet une prise en compte de phénomènes linguistiques plus riches dans des applications plus ambitieuses.

Je consacre le chapitre 3 à ces différents points.

2.2.2 Exploiter le Web comme corpus

Parallèlement à ces activités autour des corpus « classiques », je commençais à me pencher sur l'utilisation du Web comme une source de données. Le déclenchement de ces travaux fut là encore un besoin exprimé par un collègue : dans ses travaux en morpho-phonologie, Marc Plénat avait commencé à recueillir des attestations sur le Web pour alimenter ses longues listes de mots dérivés, les plus célèbres à l'époque étant les adjectifs en *-esque* (Plénat, 1997). Son approche était simple : se basant sur son intuition de linguiste il testait des dérivés possibles jusqu'ici non attestés en les tapant comme requêtes dans un moteur de recherche, et vérifiait le cas échéant qu'il s'agissait de créations lexicales authentiques (et pas des artefacts de la technique, comme nous le détaillerons dans la section 4.4). Bien que Marc Plénat soit guidé par une expertise incontestable sur la question, et doté d'un flair éprouvé, nous avions convenu rapidement qu'une grande partie de ce travail devrait pouvoir être automatisé. Il suffisait de trouver un truchement permettant d'interroger le moteur de recherche *via* un programme qui s'occuperait des tâches répétitives d'interrogation et de vérification. À l'époque, les applications en linguistique informatisée qui s'intéressaient au Web étaient très peu nombreuses, mais allaient connaître un engouement certain (c'est ce qu'indique un des nuages de la frise 0.1). Le logiciel que j'ai créé en collaboration avec Nabil Hathout, nommé (cette fois de façon transparente) Webaffix, fut un des premiers outils à exploiter le Web comme corpus linguistique, avec un objectif certes très spécifique : le repérage de nouvelles formes lexicales construites avec un suffixe donné. Je crois que je me souviendrai toute ma vie du jour où la première version du prototype fut présentée à Marc Plénat, qui voyait d'un œil humide de nouvelles attestations de dérivés en *-esque* défiler sur l'écran, abattant en quelques minutes

un travail d'orpailleur qui lui aurait demandé des semaines de fouille répétitive. Par bien des côtés cela me mettait dans la même position valorisante que lorsque j'avais montré à Jean Le Dû les grossières cartes géolinguistiques tracées automatiquement cinq années auparavant. Si ces premiers travaux fournissaient des données très bruitées qui étaient totalement acceptables pour un chercheur comme Marc Plénat aguerri au dépouillement de grandes quantités de texte, nous avions, Nabil Hathout et moi, la volonté toute talienne d'affiner le traitement et surtout le filtrage de l'outil pour lui permettre de répondre à d'autres exigences. Ce fut le début d'une longue collaboration qui s'articulait avec les compétences et les méthodes développées par Nabil en morphologie computationnelle, et qui connut plusieurs applications et plusieurs étapes, que la turbulente évolution du Web (tant technologique qu'économique) ne manqua pas d'animer sur une période de plus de huit années (Hathout et Tanguy, 2005; Hathout et Tanguy, 2002; Hathout *et al.*, 2009).

Le chapitre 4 reprend ces différents travaux en détaillant notamment les évolutions et les autres usages linguistiques du Web.

2.2.3 Identifier et exploiter la structure du discours

Plus récemment, et notamment par le biais de projets financés (comme quoi ils ont aussi l'avantage de faire collaborer des membres d'une même équipe), j'ai pu travailler avec plusieurs collègues de l'ERSS dont le centre d'intérêt est la structuration du discours. Bien que cette thématique rejoigne par plusieurs aspects les grands principes de la sémantique interprétative, notamment en plaçant le texte comme objet premier et déterminant pour l'ensemble des analyses locales, ce domaine m'était lui aussi inconnu. Si la description des structures du discours a donné lieu à l'établissement de plusieurs modèles dominants (DRT et RST notamment), les travaux initiés à l'ERSS par Marie-Paule Péry-Woodley dans le cadre du projet Annodis (Péry-Woodley *et al.*, 2009) abordent des objets textuels jusqu'ici peu étudiés, comme les structures énumératives (voir 5.1.2, page 114 pour plus de détails et des exemples). Ce projet se situe lui-même dans une lignée importante de travaux de recherche sur la structure du discours, mettant en jeu des paramètres peu exploités comme les marques visuelles qui s'articulent avec des phénomènes syntaxico-sémantiques (Péry-Woodley, 1998). Comme on aura l'occasion de le voir dans ce mémoire, ces approches en corpus sont très complexes, les objets d'étude étant à la fois difficiles à délimiter et faisant intervenir un grand nombre de dimensions langagières dans leur description. Cette complexité se traduit bien entendu directement au niveau des processus automatiques déployés (Péry-Woodley, 2005).

Le principe de l'étude à laquelle j'ai participé dans le cadre du projet Annodis est d'étudier ce type d'objet à partir des données, en commençant par une phase d'annotation manuelle sur corpus, pour en analyser dans un second temps les caractéristiques structurelles et fonctionnelles plus en détails. Cette fois par contre, le rôle de l'outil informatique est dès le début considéré comme central dans les différentes étapes. Dans la phase d'annotation en effet, les annotateurs sont aidés dans leur repérage de ces structures énumératives par un ensemble de marqueurs repérés automatiquement et qui leurs sont présentés par le biais d'une interface dédiée (GLOZZ, réalisée par les partenaires du GREYC à Caen, (Wildlöcher et Mathet, 2009)). Bien qu'aucune décision ne soit prise par l'outil, celui-ci suggère par des configurations de marqueurs denses (des adverbiaux, des démonstratifs, etc.) la possible présence d'une structure dont l'identification relève totalement de la décision humaine. J'ai donc retrouvé dans cette configuration les positions théoriques exprimées pendant ma thèse, et la volonté de limiter le rôle de l'outil, en opposant ce type d'approche à des calculs automatiques de segmentation

d'un texte comme c'est le cas dominant dans les travaux de TAL sur ces questions. Une fois les structures annotées, c'est une collection d'objets complexes qui reste à exploiter lors de la phase d'analyse. La difficulté principale provient de la complexité interne de ces objets (des segments de textes typés reliés entre eux, mais aussi des listes de marqueurs validés par les utilisateurs) et la grande variabilité (d'une paire de phrases à plusieurs sections de textes). Mon rôle fut alors d'aider au dépouillement de ces données, notamment en utilisant des méthodes outillées pour repérer des régularités et des différences entre ces centaines d'objets. Ces travaux ont permis notamment de faire émerger une typologie des structures énumératives, et d'identifier les principales caractéristiques propres à chaque type, qui confirment en grande partie les hypothèses initiales concernant la plasticité de cette forme d'organisation du discours (Ho-Dac *et al.*, 2010). Hormis les emplois classiques de mesures statistiques sur les caractéristiques de ces structures (voir le chapitre 6), j'ai dû proposer plusieurs modes de représentation permettant à mes collègues (et à moi-même) d'en visualiser les différentes aspects, et d'identifier des configurations « intéressantes ». Pour ce type de travail, rien ne vaut un schéma synthétique, aussi appauvri soit-il, par rapport à des tableaux de données.

Comme on le verra plus en détails dans la suite, ce principe de visualisation est aussi vital pour aborder la caractérisation de textes dans leur ensemble : s'il est désormais possible de repérer aisément des configurations particulières (parfois complexes) et de mesurer certaines valeurs à différents niveaux d'un texte, il est extrêmement avantageux de pouvoir visualiser globalement le fonctionnement de tel ou tel phénomène à l'échelle du texte lui-même. A plusieurs reprises, j'ai été amené à proposer des représentations graphiques de ce type, différentes à chaque fois, pour permettre l'observation de régularités (pouvant conduire à des typologies de textes) ou de lieux particuliers (nécessitant un examen local). Ce fut le cas pour l'observation des interactions entre les structures discursives (projet Annodis), la disposition des appels de citations dans un article scientifique (projet Rhecitas) ou encore l'évolution du partage de la parole entre les deux locuteurs dans une consultation médicale (projet Intermede). Le chapitre 5 présente ces approches plus en détails.

Une caractéristique importante (et croissante) des différents projets de recherche auxquels j'ai ainsi pu participer est qu'ils prennent tous en compte les méthodologies spécifiques induites par des outillages informatiques. Cette prise en compte se traduit dès la définition du travail de recherche à travers la volonté de gérer de grandes quantités de données, d'avoir une description directement opératoire des phénomènes étudiés permettant leur repérage automatique, ou de recourir à des analyses quantitatives et des modes de représentation sophistiqués. Les progressions parallèles (et la fécondation croisée) des données, des outils informatiques et des questionnements scientifiques constituent un exemple central du principe de complexification que je vais expliciter tout au long de ce mémoire.

2.3 La linguistique outillée hors les murs

Si mes premières années à l'ERSS furent exclusivement consacrées à collaborer avec mon environnement immédiat de collègues linguistes, la période plus récente m'a vu travailler avec d'autres disciplines et aborder des questions autres que la meilleure compréhension du fonctionnement du langage.

2.3.1 Terminologie et ingénierie des connaissances : extraction de structures dans les corpus

Les premiers contacts avec une autre discipline que la linguistique se firent par le biais de la collaboration de longue date entre Anne Condamines et Nathalie Aussenac-Gilles, de l'IRIT (Institut de Recherche en Informatique de Toulouse), dans le cadre de l'ingénierie des connaissances et concernent l'extraction d'informations terminologiques à partir de textes. La problématique qu'elles abordaient à l'époque était la définition et la projection (automatique) de marqueurs linguistiques permettant l'identification de relations entre termes dans des textes spécialisés (Aussenac-Gilles et Condamines, 2009). Les exemples classiques de ces marqueurs sont ceux qui traduisent une relation d'hyperonymie (*le X est un Y*), et qui peuvent présenter une grande variété formelle, variant notamment d'un genre de texte à un autre. Si à l'époque les outils permettant ce repérage se basaient sur du texte brut, nous avons vite vu (notamment à la suite des travaux avec J. Rebeyrolle sur les patrons d'énoncés définitoires) l'intérêt que présentait la même opération sur un texte étiqueté et lemmatisé. Le moteur de Yakwa fut donc en partie recyclé et adapté pour ce type de tâche, plus complexe que celle visée initialement : il s'agissait à la fois de repérer le marqueur global qui exprime la relation, mais aussi d'identifier précisément et d'extraire les deux termes reliés, ce qui pose des problèmes assez lourds en termes de définition des schémas internes de ces termes. Ce type de travaux, et l'extraction d'informations terminologiques en général, furent également le lieu d'une collaboration avec D. Bourigault, qui développait à l'époque avec C. Fabre l'analyseur syntaxique Syntex (Bourigault *et al.*, 2005) dont la fonction initiale était l'extraction terminologique, en poussant plus loin les fonctionnalités et la couverture de l'extracteur Lexter que D. Bourigault avait développé pendant sa thèse (Bourigault, 1995). C'est notamment grâce au développement de Syntex que nombre de travaux ultérieurs ont pu bénéficier d'un étiquetage syntaxique de très bonne qualité pour tout un ensemble de travaux de fouille de corpus.

Je détaille ces questions en section 3.3 (page 65), on l'on verra notamment les difficultés posées par l'exploitation des corpus annotés syntaxiquement.

2.3.2 Recherche d'information : analyse linguistique des requêtes

Une des collaborations importantes et durables que j'ai eues à Toulouse en dehors de la sphère de l'ERSS concerne les travaux que je continue à mener avec Josiane Mothe, spécialiste de recherche d'information (RI) à l'IRIT. Les points communs entre la linguistique et la recherche d'information textuelle sont évidents mais les liens sont pourtant souvent très ténus. Le matériau de base est le même que celui de la linguistique de corpus (des collections de textes numériques, avec des contenus précis à rechercher à l'intérieur de ceux-ci), les techniques de traitement des données de base sont celles du TAL robuste (normalisation des formes de surface, désambiguïsation pour l'indexation). Par contre, les pratiques sont très différentes de ce que nous faisons sur des corpus : les campagnes d'évaluation en RI et leurs mesures globales de l'efficacité sont peu stimulantes pour observer le détail du déroulement d'une chaîne de RI sur le matériau langagier. Nous avons par contre voulu profiter des nombreuses données rendues disponibles pour la communauté de RI pour regarder ce fonctionnement de plus près. L'objectif à long terme que nous nous sommes fixé est de comparer les détails des différentes méthodes d'indexation (lemmatisation versus troncation, unités lexicales simples versus complexes, filtrage ou non des mots vides, etc.) et d'appariement (entre requête et document) et d'en déduire la possibilité d'adapter ces traitements à une situation particulière,

par exemple à une requête donnée. Une des réussites les plus notables de notre collaboration (projet ARIEL) fut de mettre en évidence une corrélation faible mais significative entre les caractéristiques linguistiques des requêtes et l'efficacité d'un système de RI. Voir à ce sujet la section 6.4.2 (page 159) et (Mothe et Tanguy, 2005).

Par la suite, nous avons obtenu le financement d'un projet plus ambitieux, qui cherche à tirer profit des informations contextuelles d'une activité de recherche d'information (projet CAAS). Si la notion de contexte demande encore à être précisée, elle recouvre dans la vision actuelle les caractéristiques de la collection de documents visée, les besoins d'informations de l'utilisateur, ses connaissances et son niveau d'expertise, et les différents paramètres du moteur d'indexation et d'appariement. Mon travail dans ce cadre se concentre sur les requêtes, dans le prolongement des travaux précédents. Cette fois par contre, notre étude se base sur de grandes quantités de requêtes réelles, telles que fournies par les logs de différents moteurs de recherche existants, et non pas sur les données plus artificielles et sporadiques des campagnes d'évaluation comme TREC [1]. C'est dans ce cadre également que se déroule le projet qui finance le doctorat de S. Leva (voir section 7.3.3, page 192).

2.3.3 Autres sciences humaines et sociales : au-delà de la lexicométrie

En plus de mon environnement immédiat (le laboratoire CLLE-ERSS) et de ses collaborations académiques historiques (l'informatique à l'IRIT), ces dernières années ont vu apparaître des travaux qui ont pour vocation de répondre à un ensemble de besoins émanant des sciences humaines. Le contexte de l'université de Toulouse le Mirail est bien entendu décisif, puisqu'il regroupe en un seul établissement (et un seul bâtiment : la Maison de la Recherche) une très grande variété de disciplines, dont la plupart utilisent comme matériau de base des textes. Les besoins des SHS en termes d'exploitation de textes peuvent parfois rejoindre certains des nôtres, et les bénéfices qu'elles peuvent tirer des technologies du TAL sont potentiellement très importants. C'est le cas, dans notre expérience, des sociologues, des psychologues et des littéraires (mais il faudrait bien entendu y ajouter les historiens).

Plusieurs collaborations ponctuelles ont eu lieu à l'initiative de collègues ou d'étudiants, qui concernaient pour la plupart des besoins assez simples en termes d'exploitation automatique de petits corpus spécifiques (essentiellement littéraires). Dans ces cas-là, les outils de base de la linguistique de corpus s'avèrent amplement suffisants pour répondre à la demande : table de fréquences et concordanciers font merveille et dans certains cas évitent aux intéressés de fastidieux comptages et recueils manuels d'information autour d'une partie du lexique visée par leur étude. Dans des cas plus complexes, par exemple lorsqu'il s'agit de comparer le vocabulaire de plusieurs textes ou sous-corpus, les calculs de spécificités proposés par des logiciels comme Lexico (Lebart et Salem, 1994) remplissent très bien leur fonction. Là encore, ces demandes spontanées permettent agréablement de rendre service, et de se familiariser avec les préoccupations d'autres disciplines.

En ce qui concerne les rapprochements à l'initiative des linguistes, ils consistent essentiellement à proposer des techniques plus complexes issues de la linguistique outillée. D. Bourigault et moi-même avions ainsi décidé en 2000 de participer à une journée organisée par notre école doctorale sur les outils d'analyse de textes. Nous y avons donc présenté à des doctorants et chercheurs de sociologie et psychologie (essentiellement) notre façon de fouiller les textes : extraction de syntagmes, fréquence et distribution des unités complexes, définition et affinage

1. *Text REtrieval Conference*, qui réunit tous les ans, depuis 1992, la communauté scientifique autour de compétitions sur les différentes tâches de la recherche d'information

de patrons spécifiques, études des marqueurs de relations. L'accueil fut plutôt glacial, surtout lorsqu'à notre suite furent présentés des outils d'analyse de contenu comme Tropes et Alceste. Ces outils proposent un ensemble de caractérisations plus ou moins automatiques de textes, basées sur des méthodes purement lexicales et des scénarios bien établis. Par exemple, Alceste (Reinert, 1990) permet de découper le corpus en classant les segments de texte et le vocabulaire associé à ceux-ci par le biais d'une classification hiérarchique, et propose à la fin des classes lexicales que l'on peut croiser avec des caractéristiques externes des corpus dans le cas d'une partition préalable (par auteur, par date, par caractéristique de locuteur, etc.). Tropes (Ghiglione *et al.*, 1998) va jusqu'à proposer une analyse thématique automatique en se basant sur des dictionnaires préétablis (et modifiables), et présente les résultats en termes de zones thématiques dans un texte. Ces approches sont classiquement utilisées en sciences sociales pour dépouiller des réponses ouvertes à des enquêtes, en permettant de croiser les caractéristiques des informateurs avec les unités lexicales qu'ils utilisent, ou encore à caractériser globalement certains textes (retours d'expérience, documents techniques, discours politiques, etc.). Dans tous les cas ils prennent en charge l'ensemble des problèmes de traitement de surface du langage (découpage des unités, normalisation des formes) et proposent des classifications projetées sur des données brutes sans intervention de l'utilisateur. Le rôle de ce dernier se concentre donc dans ces approches sur l'interprétation des nombreuses représentations des résultats d'analyses, généralement en s'appuyant sur des pratiques connues en analyse quantitative des données (analyses factorielles, classifications hiérarchiques).

Il est clair que comparées à ce type de solution, nos façons d'aborder la question n'apportaient que des problèmes : nous insistions sur la complexité des phénomènes langagiers (ambiguïté, limites d'une approche purement lexicale, impact des modalisations, de la négation, etc.) et la nécessité de déployer et d'adapter des procédures minutieuses. Refroidis par cette expérience, mes collègues et moi-même avons délaissé pendant quelques temps ce type de dialogue, d'autres expériences peu enthousiasmantes venant nous confirmer que nous nous trompions peut-être tout simplement de public : certains chercheurs venaient nous voir avec des textes de très petit volume (parfois limités à quelques pages), voire manuscrits... Ce n'est que plus récemment que des collaborations fructueuses ont pu se mettre en place.

Le succès relatif de ces dernières peut être imputé à plusieurs facteurs :

- un meilleur positionnement de notre part, en intégrant mieux les besoins réels de ces disciplines et en prenant le soin de distinguer précisément les apports de notre approche en comparaison des méthodes lexicométriques. C'est le cas du projet Intermede (Vergely *et al.*, 2009; Tanguy *et al.*, 2011a) où nous avions justement comme partenaire Pascal Marchand, spécialiste des approches lexicométriques, et où nous avons dû innover pour proposer d'autres méthodes d'analyse des données (voir section 5.2.4, page 128) ;
- un rapprochement de notre mode de travail avec les méthodes utilisées pour la quantification et l'analyse statistique en SHS. Dans le projet Intermede toujours, nous avons utilisé des analyses factorielles pour étudier les relations entre les traits linguistiques et les caractéristiques des locuteurs ;
- une insatisfaction de la part des collègues vis-à-vis des méthodes sus-citées, notamment par rapport à leur incapacité à prendre en compte des phénomènes complexes liés à certains aspects de l'expression. Un travail envisagé avec le laboratoire LISST-CERS (UMR 5193) sur l'analyse des rapports d'expertise psychiatrique dans le cadre des procès d'assises va justement tenter de pallier les problèmes de la prise en compte de marques particulières de la modalité et de la subjectivité ;
- des besoins spécifiques sur des données particulières que les modèles classiques ne sont

pas destinés à traiter (dialogues dans Intermede, citations dans les articles scientifiques pour le projet RESOCIT, textes techniques structurés pour les documents professionnels, voir ci-dessous) ;
– un effort dans la présentation des résultats et des méthodes, notamment par le biais de techniques de visualisation qui, comme je l'ai dit, sont de bons vecteurs de communication interdisciplinaire (voir le chapitre 5 qui leur est consacré).

2.3.4 Entreprises : des données et des problèmes nouveaux

Depuis très longtemps les chercheurs de l'ERSS ont pris l'habitude de collaborer avec le monde « industriel », par le biais de conventions de recherche, et en produisant des efforts pour valoriser les apports de la discipline à des questions pratiques. Anne Condamines, notamment, a depuis très longtemps œuvré avec succès pour démontrer l'intérêt de techniques linguistiques dans l'exploitation des textes professionnels, par le biais de la terminologie (et ses implications dans l'indexation documentaire). Le contexte toulousain aidant, la plupart de ces applications se situe dans le périmètre de l'aéronautique, que ce soit pour la documentation technique (Airbus et EADS), la terminologie et les ontologies (CNES) ou les rapports d'incidents (CFH). Toutefois, le spectre a tendance à s'élargir au-delà de ces domaines, notamment vers des entreprises spécialisées dans la recherche et le filtrage d'information (Orange Labs, Synomia, Exalead, Atchik, etc.).

Ces collaborations ont toujours présenté plusieurs intérêts : ouvrir des problématiques nouvelles, accroître la visibilité de notre discipline (ce qui correspond de plus en plus à un besoin exprimé dans le monde académique), accéder à des données intéressantes (des textes spécialisés inaccessibles autrement) avec la possibilité d'interagir avec des experts du domaine, sans compter les retombées financières pour le laboratoire (et surtout les étudiants en thèse ou en post-doctorat). Sur le versant pédagogique, de telles collaborations ont en plus la vertu d'apporter à nos étudiants de master des contacts enrichissants et motivants avec des problématiques réelles et contemporaines.

Pour ma part, je peux citer le travail en cours avec la société CFH (Conseil en Facteurs Humains) sur un ensemble de problématiques liées à la sécurité, et plus spécifiquement à celle du transport aérien. Cette collaboration prend la suite de travaux initiés par Didier Bourigault et Cécile Fabre, qui visaient l'aide à la classification automatique de rapports d'incidents rédigés par les experts du BEA (Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile), en utilisant un extracteur de termes et un système par apprentissage pour proposer aux experts des catégories et champs de classement prédéfinis à partir du texte narratif qu'ils rédigent. Un ensemble d'applications autour de ce type de données sont actuellement en cours d'élaboration et pour certaines, déjà déployées : repérage des anomalies dans la classification (correction des erreurs), analyse des tendances dans les flux de rapports, repérage de « signaux faibles », ou signes avant-coureurs de problèmes futurs (voir section 7.3.2 pour plus de détails).

La collaboration avec CFH est une preuve (s'il en fallait) de l'intérêt de ce type d'approche : au fil des années l'activité de cette entreprise de facteurs humains s'est de plus en plus concentrée sur des problématiques linguistiques, si bien que l'essentiel de sa main d'œuvre actuelle est composée de linguistes informaticiens (pour la grande majorité issus de nos formations). Le doctorat de N. Tulechki que je co-encadre actuellement avec Marie-Paule Péry-Woodley est la consolidation de cette relation au travers d'une convention CIFRE.

2.4 Une phase de transition vers des méthodes nouvelles

Comme on le voit, la richesse et la variété des sujets de recherche, les contacts stimulants avec tous les domaines de la linguistique font de l'ERSS un lieu privilégié. En plus de 10 ans, ce laboratoire a bien entendu évolué, et la place du TAL y est plus importante qu'initialement. Le fait qu'un axe de recherche (que j'ai le plaisir de coordonner depuis 2008) y porte ce nom traduit bien cette évolution, qui a su placer le TAL à la fois comme interlocuteur de la linguistique descriptive, mais aussi comme discipline à part entière.

Les dernières années ont vu cependant le TAL évoluer en tant que discipline, et se donner de nouvelles méthodes et de nouveaux outils, qui ont bien entendu modifié nos pratiques. Le changement le plus important est sans aucun doute le développement des méthodes par apprentissage, qui sont devenues le mode opératoire principal pour aborder l'ensemble des questions soulevées par l'exploitation et l'analyse des données langagières (ce changement est indiqué par un des nuages de la frise de la figure 0.1). La première conséquence que ce changement en profondeur a eu, de mon point de vue, est le creusement d'un fossé entre les communautés de la linguistique et de l'informatique, qui jusque là avait su garder une proximité féconde. J'ai ainsi vu plusieurs de mes collègues revenir parfois découragés de conférences dans lesquelles les techniques présentées étaient difficiles d'accès et opaques, laissant peu de place aux questionnements plus linguistiques et donc aux linguistes. Nous avons donc (et sans doute moi le premier) décidé d'infléchir nos propres pratiques et tenté de trouver un positionnement confortable par rapport à cet état de fait. Cela m'a notamment conduit à expérimenter avec ces nouvelles méthodes, à me familiariser avec certaines d'entre elles, et à réfléchir à leur intégration dans les différents travaux du laboratoire. Cette transition est toujours en cours, et facilitée par les étudiants, sans doute plus réactifs et plus enthousiastes. Sur un plan plus optimiste, elle m'a permis également, comme à tant d'autres, de pouvoir traiter plus efficacement certaines parties de mon travail, notamment en mettant en place plus rapidement des systèmes opérationnels et en offrant des possibilités d'aborder des données complexes. Il convient toutefois de bien prendre la mesure des implications de ce changement de paradigme pour le TAL tel qu'il se manifeste dans un laboratoire de linguistique. J'aborde ces questions en détails dans la dernière partie de ce mémoire (chapitres 7 et 8), tant du point de vue des réalisations faisant usage de ces techniques que des réflexions plus globales qu'elles impliquent pour les rapports entre les disciplines, et la place de la linguistique dans ce nouveau paysage.

Deuxième partie

Rendre les données accessibles : les corpus et le Web

Que ce soit pour le volet applicatif du TAL, avec l'utilisation croissante de grandes quantités de données pour établir des modèles prédictifs de la langue (comme le résume le célèbre adage « *more data are better data* » de Church et Mercer (1993)), ou sur le plan de l'investigation linguistique (« *The more you can gather, the clearer and more accurate will be the picture that you get of the language* » de Sinclair (2004)), l'utilisation de données massives est devenue la règle des approches empiristes. De ce fait, une partie importante de mon activité a consisté à participer à des efforts pour rendre accessible à l'analyse des collections de textes électroniques, et à proposer des modes d'accès plus sophistiqués que la simple recherche de séquence de caractères dans un texte brut.

Cette partie regroupe peut-être les aspects les moins nobles de l'informatique, puisqu'ils ne concernent que l'*accès* aux données. Les programmes typiques dédiés à cette tâche ont généralement comme but principal de gérer la quantité, sans prétention a priori à s'intégrer à l'analyse elle-même. Nous verrons que ce n'est justement pas le cas, et que des développements spécifiques doivent être envisagés à chaque fois qu'une nouvelle problématique émerge.

Les travaux présentés ici sont aussi ceux qui semblent le plus éloignés des questions plus centralement linguistiques, car ils laissent aux utilisateurs de ces méthodes informatisées les tâches plus élevées de définir la question de recherche, et de tirer les conclusions. Mais c'est justement parce que les modes d'accès informatisés aux données sont au centre d'un grand nombre de travaux qu'il est important de s'accaparer aussi ces questionnements et de ne pas se réfugier dans la seule technique. C'est ce que j'ai essayé de faire systématiquement : en entrant dans les problématiques de plusieurs collègues, je pense avoir apporté plus qu'un simple soutien logistique ou méthodologique, et acquis une sorte de recul par rapport aux pratiques et aux enjeux.

Les travaux que j'ai menés sur cette question peuvent se décliner en deux grandes catégories. La première concerne l'exploitation de corpus « classiques », c'est-à-dire de collections de textes écrits rassemblés spécifiquement pour une étude linguistique. Que la collection soit l'objet d'une étude qui vise à en identifier des caractéristiques particulières, ou un échantillon supposé représentatif de productions langagières considérées comme génériques, les approches techniques sont les mêmes. Nous verrons donc sous l'angle des expériences passées la variété des besoins en termes d'accès à ces données.

La seconde concerne une évolution plus récente dans ce type de travail, qui cherche à tirer partie du Web et à envisager son utilisation comme source de données langagières. Si c'est bien la quantité impressionnante qui est la motivation principale, d'autres considérations positives plus qualitatives justifient les efforts toujours en cours dans la communauté du TAL pour lutter contre les difficultés inhérentes à ce matériau.

Dans les deux cas, si les techniques précises ont varié en fonction des contraintes inhérentes aux données et des informations disponibles, plusieurs principes communs peuvent néanmoins s'en dégager, et c'est ce que je propose de faire au fil de leur présentation.

Je tâcherai aussi de montrer comment, au fil du temps, les techniques ont évolué et modifié la nature des relations entre la linguistique empirique et les données sur lesquelles elle se base.

Chapitre 3

Fouiller les corpus : du texte brut aux annotations

Ce chapitre concerne les méthodes d'exploration et d'interrogation de corpus écrits, et plus spécifiquement de corpus préalablement annotés. Les travaux que je présente ici sont ceux que j'ai réalisés essentiellement lors de mes premières années à l'ERSS (de 1999 à 2002 environ) et se placent sur un terrain situé à mi-chemin entre le TAL et la linguistique de corpus : ils concernent un outillage fondamental de la linguistique empirique, peut-être le principal et en tout cas celui qui m'a permis de me familiariser avec le transfert des technologies informatiques vers les usages de la recherche sur les mécanismes du langage.

J'ai replacé dans la frise chronologique de la figure 3.1 les différentes avancées techniques concernant à la fois les procédures d'interrogation de corpus et les outils d'annotation qui ont permis de faire progresser celles-ci. On peut y voir que les premiers outils informatiques de base pour explorer des corpus sont apparus au cours des années 1960 avec les premiers concordanciers [1].

Plusieurs évolutions sont visibles sur cette frise :

- l'évolution de la **taille des corpus** : j'y ai indiqué, dans une échelle logarithmique seule à même de rendre visible l'explosion des données accessibles, quelques points de repères correspondant à des corpus génériques de l'anglais. Depuis le *Brown Corpus* des années 1960 jusqu'aux données massives accumulées par Google à travers leurs efforts de numérisation d'ouvrages, on voit bien que la progression est exponentielle.
- l'apparition des **analyseurs automatiques robustes** : les premiers analyseurs syntaxiques historiques datent de la fin des années 1950, mais ce sont sans doute les étiqueteurs morphosyntaxiques qui ont le plus influencé les travaux sur corpus, et se sont le mieux diffusés dans la communauté des linguistes descriptivistes. On peut voir qu'en ce qui concerne le français le tournant se situe essentiellement au milieu des années 1990, avec la mise à disposition (ou la diffusion commerciale) de plusieurs outils en un laps de temps très réduit. En ce qui concerne les analyseurs syntaxiques à proprement parler, ils sont arrivés (ou revenus) un peu plus tard, et je n'ai cité que celui avec lequel j'ai vraiment eu l'occasion de travailler, à savoir Syntex, et le tout nouvel analyseur syntaxique développé à l'ERSS par Assaf Urieli, Talismane (voir section 7.3.4, page 193).
- l'évolution des **outils d'interrogation de corpus** : il s'agit d'une évolution logique

1. L'outil dont j'ai pu trouver l'évocation la plus ancienne est *Cocoa*, développé en 1967 par D.B. Russell, et utilisé notamment à Oxford.

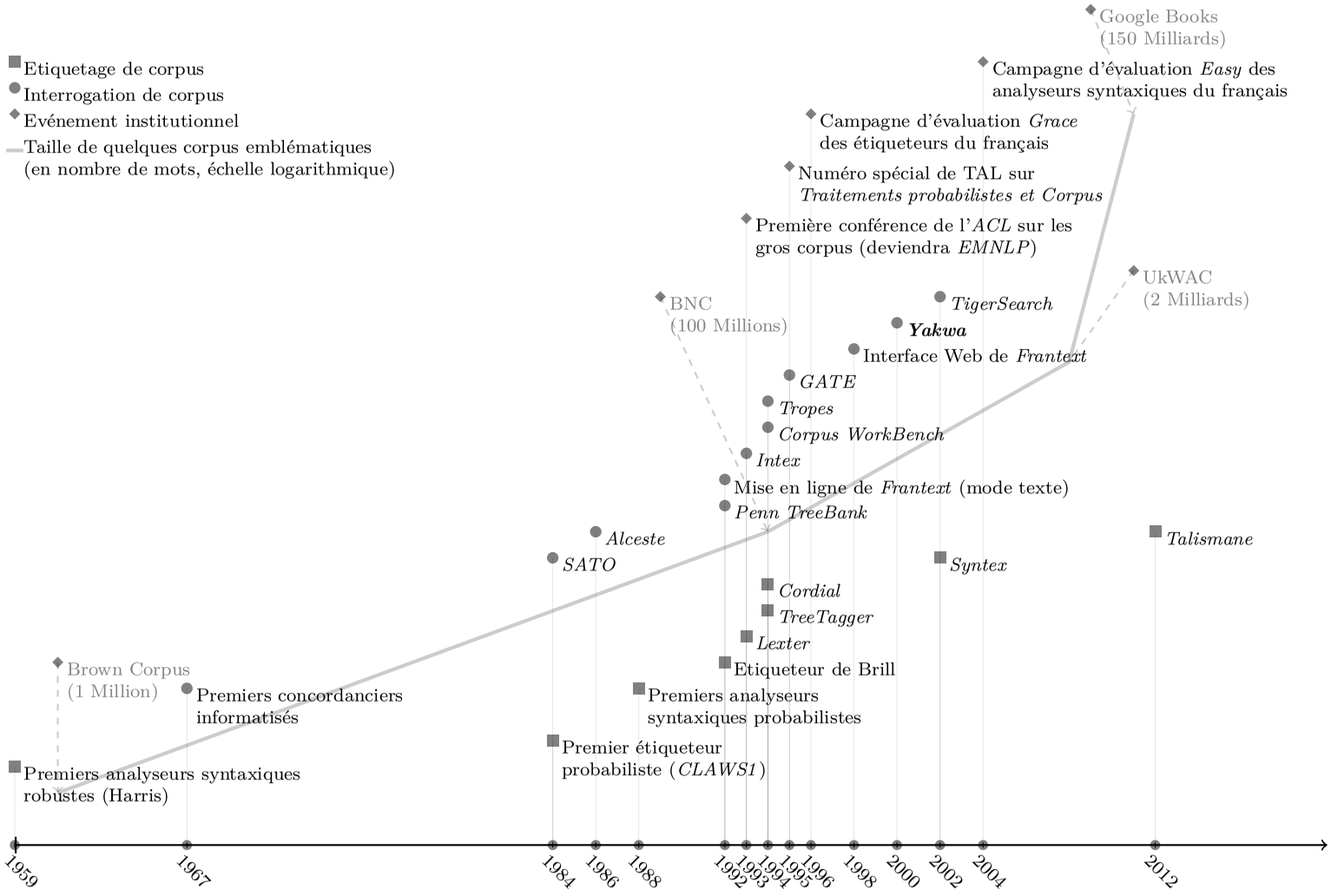

FIGURE 3.1 – Frise chronologique de l'annotation et de l'interrogation des corpus

et parallèle en réponse aux deux points précédents, puisque de nouveaux outils et de nouvelles méthodes ont dû apparaître pour faire face à la masse et aux nouvelles informations disponibles sur lesquelles les linguistes pouvaient s'appuyer pour effectuer leurs recherches dans les textes. J'ai placé sur le graphique des outils de différents types : des outils très génériques de recherche de mots ou de séquences, mais aussi des approches plus guidées d'analyse de contenu.

– l'**organisation du monde de la recherche** autour des questions soulevées par les usages et les techniques liées aux corpus. J'y ai indiqué à la fois les conférences qui prennent acte des préoccupations de la communauté du TAL, notamment autour de l'évolution de la taille des corpus, et les campagnes d'évaluation qui structurent la diffusion et la progression technique des analyseurs robustes (morphosyntaxiques, syntaxiques et autres). Ces différents événements traduisent bien à mon avis la place centrale que prennent ces questions, notamment au moment (le milieu des années 1990) où le générationalisme cède la place aux approches empiriques ce qui, dans le champ abordé dans ce chapitre, se traduit par l'effacement des analyseurs syntaxiques basés sur des grammaires au profit d'outils probabilistes.

La forme utilisée pour ce graphique (comme pour les autres de ce mémoire) tente de traduire la montée en puissance des techniques massives et les questions nouvelles qui se posent aux linguistes travaillant sur corpus. C'est ce point particulier que je vais tenter d'illustrer ici, à travers les différents travaux d'exploitation et d'outillage auxquels j'ai personnellement participé.

3.1 Inventaire des besoins et des pratiques

Je vais dans un premier temps résumer les différents besoins que j'ai pu rencontrer en termes de recherche dans des corpus électroniques, et d'en dégager une première typologie qui articule les exigences techniques et les types de pratiques.

3.1.1 Premier niveau : recherche d'attestations d'unités simples

Malgré la complexité des techniques et des ressources disponibles, comme je les présenterai succinctement dans les sections suivantes, les premiers types de besoins exprimés par des linguistes souhaitant travailler sur des données attestées sont extrêmement simples, sans grande exigence technique. Que ce soit pour étudier un élément du lexique, ou plus fréquemment dans mes expériences des mots grammaticaux divers (prépositions, conjonctions, connecteurs [2]), le but est bien souvent limité à la recherche d'occurrences afin de récolter une liste d'attestations qui seront généralement examinées minutieusement à la main par le commanditaire.

Les besoins en outillage pour ce type de travail sont assez légers. Bien souvent, un simple concordancier fait amplement l'affaire (quand ce n'est pas la fonction *rechercher* d'un logiciel de traitement de texte), et les fonctionnalités de présentation en liste synthétique des résultats sont tout à fait adaptées à la phase d'analyse ultérieure, exception faite des études à visée discursive qui nécessitent un accès à un contexte plus large que les quelques dizaines de caractères à gauche et à droite du pivot classiquement présentées dans les concordances. De

2. Les conjonctions et prépositions spatiales et temporelles ont par exemple constitué une demande assez importante à l'ERSS.

la même façon, ce type de demande visait rarement un corpus précis, et toute source était susceptible de convenir si elle fournissait de bons exemples variés.

Quitte à plonger dans la caricature, il me semble que cela correspondait simplement à une évolution de certains linguistes vers un usage progressif des données attestées, et l'on a pu assister globalement à l'apparition timide mais en progression constante d'énoncés « réels » dans les exempliers (en complément, puis en lieu et place des exemples forgés). Si certains avaient leur corpus spécifique, les grandes sources de données utilisées furent (et sont toujours) la base Frantext, les corpus journalistiques des grands quotidiens français distribués sur CD-ROM et, par la suite, le Web. A chaque fois, le mode d'accès était de toute façon fourni avec la source : interface Stella pour Frantext, outils d'interrogation dédiés distribués par ELRA [3] pour les journaux, et moteurs de recherche comme Google pour le Web. Hormis pour ce dernier cas, des fonctionnalités de base comme la recherche par troncation voire la catégorisation des unités lexicales sont facilement accessibles pour gérer les ambiguïtés et les flexions. La quantification est également un aspect important et pris en compte dans ce type de travaux, mais là encore les fréquences brutes sont fournies par l'outil d'interrogation.

L'utilisation de corpus différents (essentiellement suivant le genre ou la période) était également une demande croissante, mais émanait souvent d'un autre type de chercheur, déjà plus sensibilisé à la linguistique de corpus.

Comme je l'ai dit, le travail de dépouillement et d'analyse des attestations ainsi repérées « en vrac » par une recherche automatisée était bien souvent effectué à la main. Toutefois, il est important de préciser que des outils plus avancés proposent un ensemble de fonctionnalités et de mode opératoire permettant d'assister ce type de travail. Je reprendrai ici l'exemple de l'utilisation des concordances (appelées également *KWIC* pour *Key Word In Context*) dans la tradition anglo-saxonne de lexicographie à partir de corpus, comme présentée par Atkins et Rundell (2008).

Suivant un schéma classique (en fait une consigne de travail pour les lexicographes), la recherche commence par un repérage des différentes occurrences d'une vedette comme *taste* (*op. cit.* page 104). La flexion doit être bien entendu envisagée, soit en séparant spécifiquement les formes (*tasted, tasting*), soit en utilisant une recherche qui prend en compte le lemme des unités lexicales. L'étape suivante est le repérage de certaines régularités dans les contextes (aidé par des fonctionnalités de classement en fonction des mots situés immédiatement à gauche ou à droit du pivot), par exemple la présence d'adjectifs (*good*, *bad*, etc.) à gauche, puis le ciblage de ce type d'occurrences. Dans ce cas, il est nécessaire d'étendre le mode d'interrogation du corpus en utilisant des requêtes plus complexes, par exemple en recherchant justement des schémas couvrant plusieurs mots, et faisant intervenir des contraintes catégorielles, comme *Adj taste* ou encore *in Adj taste*.

Ce travail consiste donc en une interaction plus forte entre le chercheur et l'outil, et suppose du premier une compétence plus importante comme nous allons le voir. Ce type de pratique se rapproche donc d'un niveau plus complexe d'utilisation des outils d'interrogation de corpus, qui se distingue essentiellement par le type d'objet linguistique étudié.

3. `http://catalog.elra.info/`

3.1.2 Deuxième niveau : patrons morphosyntaxiques pour accéder aux structures

Au delà du simple mot, de nombreux travaux (majoritaires dans mon expérience) se concentrent sur des structures phrastiques. Ce type de demande est bien entendu beaucoup plus exigeante que la précédente, et l'outillage doit être sensiblement plus sophistiqué.

Comme le dit Sarah Leroy, qui a travaillé pour sa thèse sur les antonomases et s'est donc confrontée à ce type de besoin (Leroy, 2004) :

> « *Or le fait de travailler sur une structure, et non une forme, particulière de la langue complique la recherche automatique d'attestations, puisqu'elle ne permet pas la simple recherche par chaîne de caractères qu'on peut pratiquer à l'aide de n'importe quel logiciel de traitement de texte ou moteur de recherche.* »

Je me baserai ici sur deux études effectuées à l'ERSS (deux doctorats) pour lesquels ma participation a été la plus active. Je prendrai donc comme exemples le travail de thèse de Pascale Vergely sur l'expression du dysfonctionnement technique dans les dialogues entre contrôleurs aériens (Vergely, 2004) et celui de Josette Rebeyrolle sur les énoncés définitoires (Rebeyrolle, 2000). Ce type de travail combine une définition des propriétés syntaxiques des énoncés et leur caractérisation sémantique et pragmatique.

3.1.2.1 Premier exemple : l'expression du dysfonctionnement technique

Le travail de thèse de Pascale Vergely (dirigé par Andrée Borillo et Anne Condamines) est un de ces passionnants exemples où la linguistique se penche sur des phénomènes langagiers dans le monde professionnel. Dans le cadre d'une convention de recherche avec le CENA (Centre d'Etude de la Navigation Aérienne), le travail de P. Vergely consistait en l'étude des modes d'expression d'un problème technique rencontré par les contrôleurs aériens. Il s'agissait donc pour elle d'établir (entre autres) une sorte de grammaire des énoncés du type :

J'ai un digit qui est en panne
J'ai une platine téléphonique qui m'a l'air bloquée
Y a l'imprimante heu PVA qui est un peu malade
L'anti-recouvrement marche pas bien
On a des problèmes de connexion avec le STIP

Les données exploitées consistaient en un corpus recueilli spécifiquement pour l'étude par retranscription de dialogues enregistrés en situation réelle dans une salle de contrôle aérien (le travail de P. Vergely comportait également un volet comparatif, en étudiant les variations dans des contextes techniques différents). Le fait de travailler sur des données orales retranscrites et non de l'écrit pose bien entendu des problèmes spécifiques pour leur exploitation informatique, mais je ne les aborderai pas ici.

Une telle étude vise plusieurs objectifs, parmi lesquels la définition précise des variations de l'expression du phénomène ciblé, la proposition d'une typologie et comme je l'ai dit l'étude des variations entre différents contextes de production.

Du point de vue des besoins en outillage, P. Vergely a exploité son corpus en y recherchant les séquences correspondant à des patrons comme celui que l'on peut dans un premier temps formaliser ainsi :

`[(il y a/Pro.pers + avoir) + SN + PR]`

Soit en glosant : un énoncé commençant par *il y a* (ou de fait *y'a*) ou bien par un pronom personnel suivi du verbe *avoir*, puis un syntagme nominal suivi d'une relative (ce schéma correspond aux trois premiers exemples rappelés ci-dessus). Bien entendu, des contraintes supplémentaires doivent être ajoutées pour spécifier que l'énoncé concerne bien le dysfonctionnement, notamment ici dans la relative qui doit contenir en plus des marques comme des structures verbales particulières (*être en panne*, *être H.S.*, *ne pas/plus marcher/fonctionner*, etc.).

Pascale Vergely a ainsi été une des utilisatrices de Yakwa, outil avec lequel elle a défini un ensemble de patrons morphosyntaxiques cherchant à capter les schémas de ce type. Je la cite lorsqu'elle précise la portée exacte de cette utilisation :

> « *L'objectif général n'est pas en effet de proposer, dans une perspective TAL par exemple, des patrons syntaxiques de recherche automatique des expressions du dysfonctionnement technique (EDT). Il s'agit d'une étude linguistique des EDT à l'oral dont le but consiste à mettre en évidence le lien qui existe entre les structures syntaxiques et sémantiques, la présentation de l'information, et les effets pragmatiques induits dans le but d'établir par la suite les règles qui régissent l'énonciation d'un dysfonctionnement technique. L'outil demeure dans cette perspective un soutien nous permettant de rechercher et de valider de manière plus rapide les structures, mais aussi éventuellement d'élargir les résultats élaborés manuellement.* »
>
> Vergely (2004)

On va voir que le cas est légèrement différent pour la deuxième étude que je vais présenter.

3.1.2.2 Second exemple : les énoncés définitoires

Le doctorat de Josette Rebeyrolle (Rebeyrolle, 2000), également dirigé par Andrée Borillo et Anne Condamines concernait l'étude sur un corpus varié des énoncés qui contiennent la définition d'un terme d'un langage de spécialité. Voici quelques exemples :

- *La vase peu colonisée, recouverte plusieurs heures à chaque marée, se nomme une* ***slikke****.*
- *L'expert devait se remémorer l'incident et disposait des informations contenues dans la* ***main courante*** *(journal de bord où sont notées toutes les interventions sur le réseau de distribution de gaz).*
- *Un* ***jeu de barres*** *est un circuit triphasé auquel peuvent être raccordés tous les départs (lignes, transformateurs) à une même tension.*
- *On donne le nom de* ***boutonnière****, ou* ***bray****, aux dépressions allongées, évidées dans les formations peu résistantes des séries sédimentaires ployées en ondulations anticlinales peu marquées.*

Le corpus utilisé était constitué de textes de différents types (manuels pédagogiques, articles d'encyclopédie, articles scientifiques, guides techniques), et l'étude visait notamment la comparaison de fonctionnement de ces différents genres textuels. Mais la phase principale consistait, comme pour P. Vergely, en l'élaboration d'une grammaire et d'une typologie de ces énoncés, qui passait donc par l'établissement d'une série de patrons, comme ceux exprimés ci-dessous :

- énoncés de désignation : `A` $\mathtt{V}_{désigner}$ `B - X`

– énoncés de dénomination : `B - X` $V_{s'appeler}$ `A`
– énoncés de signification : `A` $V_{signifier}$ `B - B`
– énoncés introduits par *c'est-à-dire* : `A, c'est-à-dire B - X` ou `B - X, c'est-à-dire A`
– etc.

On voit là encore que ces schémas correspondent à des structures syntaxiques spécifiées par des classes lexicales (ici des classes de verbes, $V_{s'appeler}$ recouvre par exemple *s'appeler*, *porter/mériter le nom de*, voir Rebeyrolle et Tanguy (2000) pour plus de détails). Ces patrons doivent ensuite être précisés et exprimés dans un formalisme qui rend possible leur repérage automatique ; ici encore cela a été effectué en utilisant l'outil Yakwa qui sera détaillé en § 3.2.

En plus de l'intérêt pour la description d'un tel phénomène d'affiner et de préciser les patrons (comme l'a fait P. Vergely), l'expression formelle de ceux-ci peut ensuite être utilisée pour sa quantification. J. Rebeyrolle a ainsi comparé la fréquence de chaque type d'énoncés dans des corpus différents, et démontré l'affinité de certains genres textuels avec certains types de définition.

La formalisation et l'opérationnalisation du mode d'accès à ces énoncés peuvent également être utilisées pour des applications pratiques en ingénierie linguistique et en ingénierie des connaissances.

3.1.3 Troisième niveau : patrons morphosyntaxiques pour des applications

Le troisième niveau que je souhaite détailler ici se situe dans le prolongement direct de l'exemple précédent, et rentre dans le cadre des travaux qui concernent l'extraction d'information à partir de corpus. L'objectif ici est de définir des patrons correspondant à des structures qui expriment une relation sémantique précise entre des unités lexicales ou terminologiques. De nombreux travaux menés à l'ERSS et à l'IRIT, comme ceux présentés dans Aussenac-Gilles et Condamines (2009) se situent dans cette interaction entre la linguistique de corpus et l'ingénierie des connaissances.

Par exemple, la structuration de ressources terminologiques ou ontologiques peut s'appuyer sur des expressions en corpus de relations comme l'hyperonymie par le biais de différents marqueurs comme

`divers X comme Y`

correspondant à des énoncés du type :

Présent naturellement dans divers ***tissus du corps****, comme* ***la peau*** *et* ***le cartilage****, l'acide hyaluronique est utilisé...*

Le repérage de ce type de structure permet d'identifier la relation hyperonymique reliant *peau* à *tissu du corps*. L'automatisation de ces repérages peut être envisagée si l'on définit des patrons systématiques, ce que fait une discipline du TAL comme l'extraction d'information (Poibeau, 2003).

Les énoncés définitoires de J. Rebeyrolle sont un exemple parlant de ce type de patrons. Au-delà de leur intérêt descriptif, ils ont ainsi été utilisés par la suite dans un outil d'extraction automatique (ainsi qu'une partie du logiciel Yakwa) comme Caméléon (Séguéla et Aussenac-Gilles, 1999; Jacques et Aussenac-Gilles, 2006).

Dans ce type d'utilisation des outils d'interrogation, tout comme dans le précédent, le travail consiste en de nombreux aller-retours entre la mise au point des patrons et leur projection sur corpus, dans ce que Jacquemin (1997) nomme une mise au point expérimentale.

Par contre, les applications destinées à effectuer un déploiement systématique des patrons ont un niveau maximal d'exigence : la finesse des descriptions doit viser un repérage le plus précis possible, évitant plus le bruit que le silence. Dans ce cas, l'utilisation des outils d'interrogation forme un atelier de fabrication de ces patrons, et le corpus utilisé est lui-même un terrain d'expérimentation plus qu'un environnement à explorer (les patrons étant destinés à être projetés sur d'autres corpus du même type). Il est clair alors que le niveau d'exigence technique est maximal pour ce type d'application, qui se situe néanmoins dans le prolongement des travaux des deux types précédents. Il n'est par contre pas certain que les contraintes supplémentaires nécessaires (notamment contextuelles) pour renforcer l'efficacité des patrons constituent en tant que telles un résultat linguistique pertinent pour la description du phénomène.

Je vais maintenant aborder les aspects plus techniques de ces outils d'interrogation, en les distinguant en fonction des annotations qui sont appliquées aux corpus.

3.2 Interrogation de corpus étiquetés morpho-syntaxiquement : puissance d'expression et prix à payer

Le fait de disposer de corpus étiquetés morphosyntaxiquement a été une évolution notable dans le domaine de l'exploitation des corpus. La mise à disposition d'outils d'étiquetage robustes pour différentes langues (le français, comme d'habitude, étant en retard sur l'anglais) a peu à peu fait de ce type de traitement une étape fondamentale dans l'exploitation des textes électroniques en TAL. Par contre, les informations supplémentaires ajoutées aux corpus nécessitent un outillage spécifique pour leur exploitation, et des pratiques plus complexes par les utilisateurs. Je présente ici les principes et les contraintes de ce type de techniques, ainsi que ma propre expérience en terme de développement d'outils.

3.2.1 Les étiqueteurs morphosyntaxiques : des outils bien répandus

Le premier outil de ce type que j'ai utilisé fut l'étiqueteur Tatoo développé à l'ISSCO dans le cadre du projet Multext (Armstrong *et al.*, 1995). À peu près au même moment, l'étiqueteur par transformation d'Eric Brill (Brill, 1992) était adapté au français par Josette Lecomte (Lecomte, 1998) et les premières versions de TreeTagger étaient également disponibles (Schmid, 1994). Sur le plan des outils commerciaux, citons l'analyseur Cordial de la société Synapse Développements qui effectue une analyse syntaxique mais propose des sorties du même type que les outils précédents. On peut remarquer sur la frise de la figure 3.1 l'étonnant resserrement de la période temporelle durant laquelle ces outils sont apparus [4]. On notera aussi la rapidité avec laquelle la communauté s'est préoccupée de l'évaluation de ces outils, en organisant des campagnes suivant le modèle classique des compétitions organisées pour la l'extraction et la recherche d'information, notamment via l'action Grace (Adda *et al.*, 1999).

Tous ces outils remplissent une fonction précise : ils segmentent le texte en mots et signes de ponctuation, et attribuent à chaque mot sa catégorie morpho-syntaxique, avec plus ou moins

4. Ceci dit, des outils plus anciens étaient développés pour le français, notamment à l'Inalf pour l'étiquetage du corpus Frantext, mais n'étaient pas distribués.

de détails concernant les traits morphologiques, et pour la plupart en identifiant sur cette base la forme de citation (lemme) du mot désambiguïsé. Ces deux informations supplémentaires ouvrent donc plusieurs possibilités pour exploiter les textes en nécessitant toutefois de prendre en compte le découpage initial en mots.

Chaque outil utilise un format spécifique pour présenter les résultats de l'analyse, mais le plus classique est d'utiliser un format tabulé, dans lequel chaque mot est décrit sur une ligne composé de colonnes, chacune d'elles indiquant une des trois informations (forme, catégorie, lemme). Par exemple, l'extrait suivant est produit par le logiciel TreeTagger pour le français :

```
Cet       PRO:DEM   ce
extrait   NOM       extrait
donne     VER:pres  donner
un        DET:ART   un
petit     ADJ       petit
exemple   NOM       exemple
de        PRP       de
texte     NOM       texte
étiqueté  VER:pper  étiqueter
.         SENT      .
```

La première colonne contient la forme de surface telle qu'elle est trouvée (exactement) dans le texte initial, la deuxième contient l'étiquette morphosyntaxique (ici le jeu d'étiquette utilisé est la version simple fournie par défaut par TreeTagger pour le français) et la troisième le lemme du mot. Les signes de ponctuation sont, comme on le voit, considérés comme des mots, et possèdent une catégorie et un lemme.

3.2.2 Outils d'interrogation : quelques exemples

Exploiter de telles données nécessite donc de faire appel à des outils ou programmes spécialisés (Habert (2006) parlerait plutôt d'instruments) puisqu'il n'est plus possible d'utiliser des fonctionnalités simples de recherche dans un texte brut (proposées par des éditeurs ou traitement de textes). Plusieurs outils de ce type ont été proposés à la communauté, parmi lesquels je citerai :

- Yakwa : que j'ai personnellement développé comme indiqué dans la première partie de ce mémoire à la suite du projet DiET ;
- CQP (Corpus Query Processor), développé par l'IMS de l'université de Stuttgard (Christ, 1994) ;
- Stella, le moteur de recherche de la banque de textes Frantext (Bernard *et al.*, 2002). Bien qu'il ne soit pas disponible pour interroger un corpus quelconque, mais spécifiquement dédié à l'interrogation de la version catégorisée de Frantext, il remplit exactement les mêmes fonctions que les précédents. De plus, il est bien connu de la communauté des linguistes francophones et couramment enseigné et utilisé.

Tous ces outils ont la particularité d'être conçus pour exploiter des données volumineuses (selon les époques cela varie de plusieurs centaines de milliers de mots à des dizaines de millions), et sont donc généralement des instruments complexes, faisant appel à des techniques d'indexation préalable. L'inconvénient majeur (en plus du coût élevé de développement) est la complexité de la procédure d'installation et de configuration (CQP comme Yakwa fonctionnent en réseau et nécessitent un serveur Unix et un système de gestion de bases de données, et Stella n'est utilisable que sur la base Frantext), et surtout de l'injection de nouveaux textes.

De fait, il était généralement obligatoire d'avoir un informaticien sous la main pour utiliser ces outils.

Pire encore, ces outils éloignent considérablement l'utilisateur des corpus sur lesquels il se penche, rendant par exemple impossible la correction ou l'ajout d'information au cours de la recherche (c'était une fonctionnalité fort appréciée du logiciel SATO, qui permettait d'attribuer des étiquettes arbitraires aux unités du texte). Ces handicaps majeurs (mis à part pour Frantext, qui est accessible en ligne) ont largement réduit leur potentiel d'utilisation dans la communauté, qui leur préférait généralement des outils légers comme les concordanciers, malgré leur incapacité à manipuler de gros volumes et à effectuer des recherches complexes.

Quoi qu'il en soit, ces outils apportent à leur utilisateur un ensemble de fonctionnalités très importantes. En ce qui concerne les mots simples, il est donc désormais possible de les identifier par leur lemme, leur catégorie et non plus uniquement par leur forme de surface. La plupart des outils d'interrogation de corpus conçus pour exploiter de telles annotations proposent généralement toutes les combinaisons possibles.

3.2.3 Modes d'interrogation : multiplicité des langages de requêtes

La complexité de l'interrogation provient toutefois de la prise en compte de données structurées et du fait que les mots d'un texte sont désormais des unités complexes. Un outil d'interrogation de corpus étiquetés doit donc proposer un mode d'interrogation bien plus complexe que ce que permet une simple expression régulière sur un texte brut.

Je vais prendre comme exemple la recherche d'une séquence du type « *(donner OU présenter) un ADJ* exemple* », plus précisément une occurrence (quelle que soit sa flexion) du verbe *donner* ou *présenter*, suivi de *un*, suivi d'un nombre quelconque d'adjectifs (y compris aucun), suivi du mot *exemple*. Ce type de schéma est très classique, et somme toute assez simple à exprimer. Formellement, il s'agit d'une famille de séquence infinie (il n'y a pas de limite au nombre d'adjectifs précédant *exemple*) mais qui se formalisme très bien par un langage régulier, et donc par un automate à états finis. Toutefois, hormis le formalisme des outils de la famille Intex (Silberztein, 1993) comme Nooj ou Unitex qui proposent une interface de saisie en dessinant des automates, la plupart des outils d'interrogation utilisent une représentation sous la forme d'une requête exprimée par une expression spécifique comme le montrent les exemples suivants correspondants aux trois outils précités.

Yakwa : `">donner|présenter" "un" ADJ* "exemple"`

CQP : `[lemma="donner|présenter"] [word="un"] [pos="ADJ"]* [word="exemple"]`

Stella : `&c(donner|présenter) un &*&e(g=A) exemple`

Il existe bien entendu un grand nombre de variations équivalentes pour chacune de ces requêtes, qu'elles soient dues au formalisme utilisé ou au niveau de spécification. Par exemple, il est possible de préciser dans chaque cas que « *exemple* » est un nom, mais l'absence d'ambiguïté catégorielle de cette forme rend cette spécification superflue. De même, alors que « *un* » est ambigu, le contexte droit immédiat rend également inutile d'exprimer que l'on cherche ici un déterminant et non pas un pronom.

Comme on le voit, les variations formelles entre les différents langages sont assez limitées. Sans rentrer dans les détails de ces langages de requêtes on citera les principales fonctionnalités qui sont prises en compte par un tel outil (et pour lesquelles un mode d'expression existe dans le langage d'interrogation) :

– Expression des contraintes portant sur un mot, et succession de celles-ci pour former une séquence (généralement, des formules séparées par des espaces) ;

– Distinction entre la recherche sur le lemme, la forme ou la catégorie grammaticale (indication utilisant des guillemets, ou des notations spécifiques) ;
– Patrons de chaînes pour les formes, lemmes ou étiquettes des catégories (généralement par des expressions régulières) ;
– Opérations logiques (conjonction, disjonction ou négation de contraintes) ;
– Fermetures (répétition, optionnalité d'une partie du schéma).

On voit donc qu'apprendre à créer une telle requête demande un ensemble de compétences. La première, et la plus importante, est bien entendu la capacité à formaliser le patron lui-même (elle correspond à la capacité à énoncer les séquences d'éléments visés). La seconde consiste à maîtriser la syntaxe de la requête à soumettre au moteur (ce qui présuppose généralement la connaissance d'autres formalismes comme les expressions régulières), et la troisième consiste à maîtriser la liste des étiquettes des catégories grammaticales (qui peuvent varier pour un même outil en fonction de la langue ou de l'étiqueteur utilisé en amont). Ces différentes contraintes m'avaient conduit à l'époque à dépenser une grande quantité d'énergie (autant que pour construire et optimiser le moteur de recherche lui-même) pour doter Yakwa d'une interface graphique qui permettrait d'alléger cette construction. L'interface de saisie d'une requête est visible dans la figure 3.2.

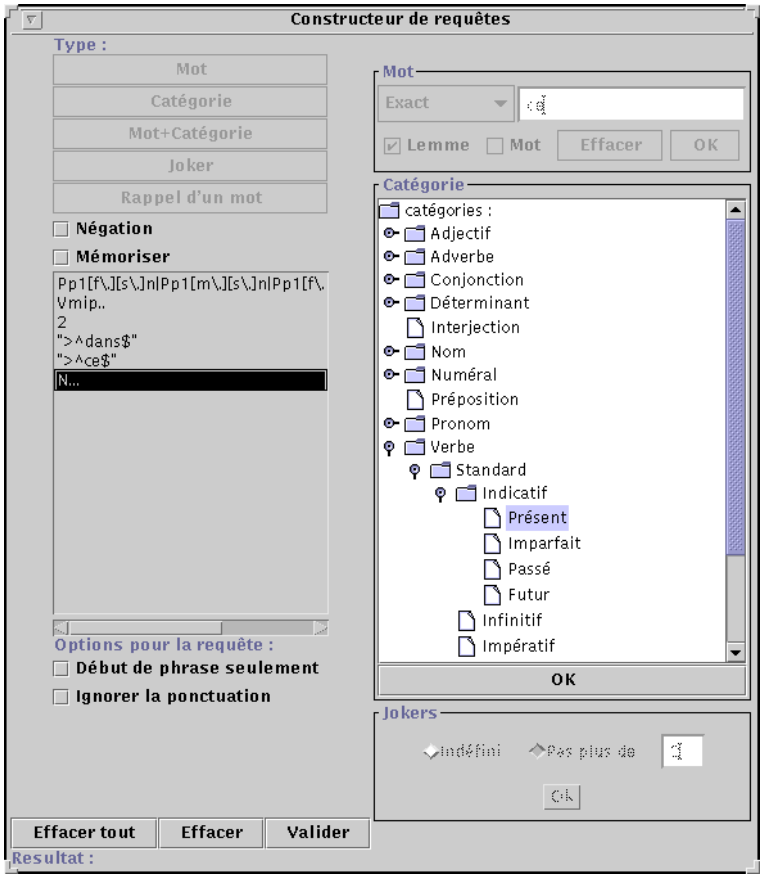

FIGURE 3.2 – Interface de construction de requêtes dans Yakwa

Comme on le voit, les différentes fonctionnalités présentées ci-dessus sont accessibles par

des boutons dédiés (« joker » représente les fermetures), et la zone centrale de la fenêtre correspond à la taxonomie du jeu d'étiquettes, qui permet ainsi une spécification à grain variable des différentes sous-catégories et traits morphologiques. La logique de construction du patron reste celle de la constitution d'une séquence de contraintes ou de fermetures, dont la trace dans le langage de requêtes apparaît à gauche au fur et à mesure.

Je peux dire avec le recul de plusieurs années, et à propos d'un logiciel que j'ai depuis longtemps arrêté de maintenir par manque de temps, qu'il a rempli le rôle qui lui avait été attribué. La plupart de mes collègues-cobayes semblaient avoir réussi à s'approprier les fonctionnalités de base permettant la définition de patrons du type de l'exemple ci-dessus.

Toutefois, et c'est là un point important sur lequel je reviendrai, cet outil permettait également des requêtes bien plus complexes, et ce sur deux dimensions distinctes.

3.2.4 Complexification des requêtes

Le premier type de complexification correspond au degré de sophistication des patrons eux-mêmes. Lors de mon travail avec Josette Rebeyrolle (Rebeyrolle et Tanguy, 2000), la définition de patrons très complexes était non seulement décisive, mais elle était surtout évaluable grâce au travail accompli par J. Rebeyrolle qui avait relevé tous les énoncés de son corpus en s'aidant du logiciel SATO pour s'appuyer sur les marques lexicales les plus discriminantes, comme les verbes de dénomination. Nous avons donc cherché à définir les meilleurs patrons possibles pour chaque type d'énoncé définitoire. Certains de ceux-ci atteignaient des niveaux élevés de complexité, comme le suivant (la syntaxe exacte en a été simplifiée par rapport au formalisme présenté plus haut) :

```
Nom|Pro (Non Vbe)* "définir" (Non Vbe)* Nom (Non Vbe)* comme (Non Adv,Pro,Prép)
```

Ce patron correspond à des énoncés comme celui-ci (la partie exacte en correspondance avec le patron est soulignée) :

Nous définissons le système d'information comme l'ensemble des moyens de traduction et d'utilisation de connaissances.

Plus précisément, ce patron recherche un nom ou un pronom, puis plus loin une forme du verbe *définir*, puis un nom, puis *comme*. Les éléments dotés de fermeture (indiqués par des étoiles dans le patron) s'appliquent à des interdictions de certaines catégories : absence de verbes entre les quatre éléments principaux du patron, et absence d'un adverbe, pronom ou préposition à droite de *comme*. Ces contraintes ont simplement pour but de filtrer des énoncés « bruitants », notamment en autorisant le patron à s'étaler sur plusieurs propositions (l'interdiction des verbes) et en évitant les autres rôles de *comme* (comme dans *comme d'habitude*, *comme il se doit*, etc.).

Dans le même ordre d'idée, il a été nécessaire de préciser le patron correspondant aux énoncés *Par X on entend Y* par :

```
par (Non Vbe (sauf modaux))* "entendre"
```

en autorisant les verbes modaux entre *par* et *entendre*, au vu d'énoncés du type *par X il* ***faut*** *entendre Y*.

La définition de ce type de patrons nécessite un ensemble d'aller-retours entre les résultats et la définition des patrons par approches successives. De ce fait, le rôle de l'outil dépasse le simple repérage d'énoncés cibles : il devient en quelque sorte un assistant à la définition d'un objet d'étude. C'est sans doute une des vertus les plus souvent passées sous silence de l'utilisation des outils informatiques, en ce sens que la rigueur qu'ils imposent peut apporter en retour une meilleur compréhension de l'objet visé.

Le second type de complexification est issu de demandes spécifiques qui ont donné lieu à l'ajout de fonctionnalités supplémentaires à l'outil. Parmi celles-ci, citons :

- la possibilité d'ignorer les signes de ponctuation : bien souvent ces éléments ne sont pas pertinents pour certains patrons, comme les éventuelles virgules précédant une relative ;
- la restriction de la position du patron en initiale de phrase : plusieurs études, notamment sur les introducteurs de cadre nécessitent ce type de contrainte, qui font partie de la définition du phénomène visé ;
- la mémorisation d'un mot et la recherche de reprise : similaire au mécanisme de mémorisation de sous-chaînes des expressions régulières, cette fonctionnalité pointue permet de rechercher la répétition d'une unité lexicale dans un patron généralement transphrastique. Cette fonctionnalité a notamment été utilisée pour affiner certains patrons trop génériques pour les énoncés définitoires, en exigeant une première mention du nom pivot. C'est le cas d'un énoncé comme le suivant, où au patron *X est un Y* s'ajoute la nécessité d'avoir une occurrence de *X* dans le contexte gauche :
  Une section se reconnaît par le fait qu'elle est composée de blocs de texte (les paragraphes) situés sous un ***titre****. Le* ***titre*** *est un bloc court isolé et typographié dans une fonte plus grasse, éventuellement souligné et numéroté.*

Malgré ces fonctionnalités ajoutées, il était clair dès le début que la limite de l'outil est rapidement atteinte lors de la définition de patrons syntaxiques complexes : que ce soit la nécessité d'autoriser au sein d'une structure des syntagmes ou propositions (incises, appositions), ou le recours au truchement d'interdiction des verbes pour garantir une relation sujet ou objet, tout cela était voué à n'être qu'une approche grossière des complexités syntaxiques rencontrées en corpus. L'étape suivante est donc une évolution naturelle, rendue possible par le développement d'analyseurs syntaxiques robustes capables d'ajouter de nouvelles informations exploitables aux corpus étudiés.

3.3 Corpus analysés syntaxiquement : un niveau supplémentaire coûteux

Dès l'année 2001, l'ERSS avait la grande chance d'avoir à sa disposition un analyseur syntaxique opérationnel, nommé Syntex (Bourigault *et al.*, 2005; Bourigault, 2007), conçu par Didier Bourigault. Les analyseurs syntaxiques du français étaient, contrairement aux étiqueteurs morphosyntaxiques, beaucoup plus rares, et surtout très peu robustes (ceux qui l'étaient étaient très difficilement accessibles). Le développement de Syntex en tant qu'analyseur en dépendances avec une architecture légère répondait initialement à des besoins en extraction terminologique et fut donc très rapidement utilisé pour extraire des syntagmes nominaux et verbaux et, *via* une analyse distributionnelle, pour extraire des relations sémantico-lexicales (Fabre et Bourigault, 2006).

Je vais commencer par en présenter les principes, en me limitant à la présentation des annotations ainsi produites, avant d'aborder la question de la difficulté de leur exploitation pour l'interrogation de corpus.

3.3.1 L'analyseur syntaxique Syntex

Syntex fonctionne en aval d'un étiquetage morphosyntaxique (réalisé par une version modifiée de TreeTagger), auquel il délègue le découpage en mots et l'attribution des catégories

morphosyntaxiques. Par un ensemble d'heuristiques appliquées en passes successives, il identifie les relations de dépendances syntaxiques et produit un graphe (parfois partiel) comme celui indiqué en figure 3.3 pour la phrase « *Il convient de le faire pour deux raisons principales.* » :

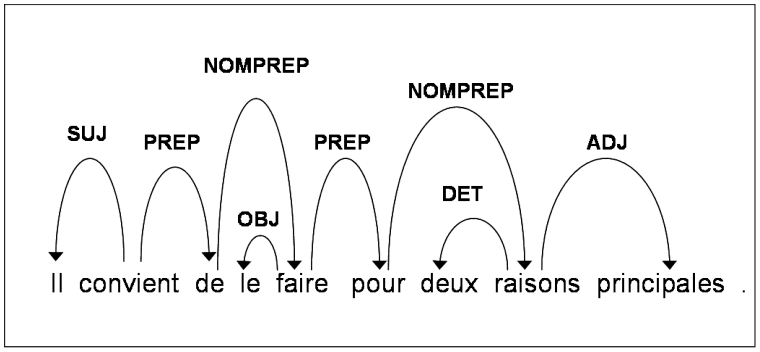

FIGURE 3.3 – Graphe de dépendances construit par Syntex

Les flèches indiquent la relation qui relie un mot dépendant (ou régi suivant l'ancienne terminologie utilisée par D. Bourigault) à son gouverneur (ou recteur). Chaque relation est catégorisée et dotée d'une étiquette indiquant la nature de la liaison syntaxique (SUJ = relation sujet, OBJ = relation objet, PREP et NOMPREP = rattachement de la préposition, etc.).

Le format de sortie de l'analyseur est bien plus complexe que celui d'un étiqueteur morphosyntaxique, mais en reprend les mêmes principes. Ci dessous se trouve l'analyse produite au format *anasynt* par Syntex pour la phrase précédente.

```
Pro|il|Il|1|SUJ;2|
VONCJ|convenir|convient|2||SUJ;1,PREP;3
Prep|de|de|3|PREP;2|NOMPREP;5 Pro|le|le|4|OBJ;5|
VINF|faire|faire|5|NOMPREP;3|OBJ;4,PREP;6
Prep|pour|pour|6|PREP;5|NOMPREP;8
Det|deux|deux|7|DET;8|
Nom|raison|raisons|8|NOMPREP;6|DET;7,ADJ;9
Adj|principal|principales|9|ADJ;8|
Typo|.|.|10||
```

Chaque mot y est décrit par l'ensemble des informations suivantes dans l'ordre, séparées par des barres verticales : sa catégorie, son lemme, sa forme de surface, son numéro d'ordre dans la phrase, la relation qui le relie à son gouverneur (s'il existe) et le numéro de celui-ci, la liste des relations qui le relient à ses dépendants (s'ils existent) et les numéros de ceux-ci.

Jusqu'ici, les utilisations des données produites par Syntex se faisaient exclusivement par des programmes dédiés (la plupart d'entre eux écrits par D. Bourigault lui-même), que ce soit pour en extraire la liste des syntagmes (et donc répondre à l'objectif initial de Syntex comme outil de terminologie) ou pour effectuer une analyse distributionnelle (voir (Bourigault, 2007) pour un inventaire des applications). Il était toutefois très tentant de proposer de telles données afin d'accroître le pouvoir d'expression d'un outil d'interrogation de corpus comme Yakwa. Mais dès lors plusieurs questions se posèrent :

– Quel moteur d'indexation et de recherche utiliser face à des données aussi complexes ?
– Quel mode d'interrogation proposer aux utilisateurs ?

– L'utilisation des informations syntaxiques allait-elle vraiment permettre d'améliorer le travail sur corpus ?

Il est tout à fait envisageable de s'appuyer directement sur les relations de dépendances pour effectuer des recherches dans les textes, comme cela a été proposé notamment par Falaise *et al.* (2011) pour les requêtes « guidées » sur le corpus Scientext ou encore par Franck Sajous à l'ERSS pour l'outil INCAS [5]. Dans ce cas, l'utilisateur doit exprimer les mots qu'il recherche ainsi que les relations syntaxiques de dépendances qu'ils entretiennent. Mais ce type de requête reste très limité et nécessite une connaissance exacte de la structure recherchée. Par exemple, la recherche d'un syntagme nominal complexe (contenant au moins deux noms) nécessite d'expliciter toutes les relations structurelles possibles entre ces noms, y compris en précisant les prépositions impliquées. Bien qu'utiles pour des cas simples comme les relations sujet ou objet, il était nécessaire de trouver un autre moyen d'exploiter ces corpus pour y rechercher des structures.

D. Bourigault, C. Fabre et moi-même avons eu la chance de nous poser ces questions au moment où l'Institut de Linguistique Française (ILF) dont l'ERSS faisait partie lançait un appel à projet pour inciter les différents laboratoires membres à collaborer sur des thématiques ouvertes. Nous avons donc proposé et obtenu un financement pour ce travail (projet baptisé *Yakwa++*), en collaboration avec nos collègues du laboratoire BCL (Bases, Corpus et Langues, CNRS et Université de Nice) : Sylvie Mellet, Etienne Brunet, Damon Mayaffre et Michèle Oliveri. Leur expérience dans l'exploitation de corpus en faisaient pour nous des utilisateurs modèles, et allait également diversifier l'éventail des besoins identifiés.

3.3.2 Outils d'exploration de corpus annotés syntaxiquement

Bien entendu le fait que nous disposions désormais d'une collection de corpus annotés syntaxiquement ne signifiait pas que de nombreux chercheurs et ingénieurs n'avaient pas été confrontés avant nous aux questions liées à leur exploitation. De nombreux corpus annotés syntaxiquement existaient déjà (même si la plupart ne concernaient pas le français), ainsi que des outils d'interrogation. Je décidai donc d'effectuer un tour d'horizon des produits disponibles avant de me lancer dans toute réalisation.

Bien que les analyseurs en dépendances soient une invention ancienne, la quasi-totalité des corpus annotés syntaxiquement sont des corpus arborés ou *treebanks*, c'est-à-dire des collections de textes dans lesquelles l'analyse a été représentée sous la forme d'un arbre syntaxique. Le plus connu est sans doute le Penn TreeBank (Marcus *et al.*, 1993), qui a inspiré des efforts similaires dans d'autres langues que l'anglais (dont le French TreeBank d'Abeillé *et al.* (2003)).

Le format natif de ces données est un format parenthésé : l'analyse syntaxique y est exprimée sous la forme de constituants imbriqués, et le format final n'est pas sans rappeler les programmes Lisp qui firent les grandes heures de l'IA. Par exemple, voici (à peu près) à quoi ressemblerait dans ce format la phrase prise en exemple précédemment (*Il convient de le faire pour deux raisons principales.* :

```
(S (NP (Pro Il) )
   (VP (V convient)
       (PP (Prep de)
            (NP (Pro le) )
```

5. Pas encore diffusé à ce jour.

```
            (VP (V faire)
                (PP (Prep pour)
                    (NP (Det deux) (N raisons) (Adj principales) )
                )
            )
        )
    )
)
```

Comme on le voit, chaque constituant est décrit par une séquence parenthésée à l'intérieur de laquelle on trouve le type du constituant (en utilisant la notation anglaise : S pour phrase, NP pour syntagme nominal, etc.) et sa composition. Dans certaines versions les rôles syntaxiques sont également exprimés (sujet, objet, etc.).

Plus tard, avec le développement de la norme XML pour le codage d'informations structurées complexes, c'est en utilisant ce langage de balises que les données ont été distribuées. Par exemple, le French TreeBank utilise un format de ce type :

```
<SENT nb="1">
      <NP fct="SUJ">
         <w cat="P" lemma="il">Il</w>
      </NP>
      <VN>
         <w cat="V"  lemma="convenir" >convient</w>
         <VP-inf fct="OBJ">
             <w cat="P" lemma="de">de</w>
             <NP fct="OBJ">
                 <w cat="P" lemma="le">le</w>
             </NP>
             <w cat="V" lemma= "faire">faire</w>
...
```

Ce format présente l'avantage majeur de comprendre l'ensemble des informations disponibles pour chaque mot (catégorie, lemme) en plus de la fonction de chaque constituant.

Plusieurs outils d'exploration de ces corpus étaient disponibles (et de nombreux ont été développés depuis), et tous se basent naturellement sur le principe de la recherche dans une structure arborée. Les plus connus sont sans doute ceux qui ont été développés spécifiquement pour le Penn TreeBank : Tgrep et Tgrep2 (Rohde, 2005). Le mode d'interrogation général est le suivant : on liste les éléments que l'on souhaite trouver dans une phrase, (des mots, ou des constituants), ainsi que les relations hiérarchiques ou positionnelles que ceux-ci doivent entretenir (X est inclus dans Y, X est au début de Y, X est à gauche de Y). L'identification de chaque élément peut faire appel au contenu textuel, à la catégorie, à une combinaison des deux, en utilisant les opérateurs habituels (expressions régulières, connecteurs logiques, etc.). Par exemple, pour trouver un syntagme nominal qui contient le nom 'raison', on emploiera une requête du type : `NP < raison`

Hormis le choix malheureux qui fait que le symbole $<$ exprime la dominance structurelle (on se serait attendu à trouver $>$), les choses deviennent rapidement plus complexes, notamment lorsque plus de deux éléments sont impliqués dans la requête.

Un autre type d'outil d'interrogation est celui qui a accompagné l'XMLisation des données. La technologie XML est en effet très bien dotée en langages de manipulation et d'interrogation des données structurées, comme les langages Xpath et Xquery. Des versions spécifiques de ces

langages ont donc été proposées pour faciliter la recherche dans ce type de données, comme l'ont fait Bird *et al.* (2006). Dans ce cas, rechercher un syntagme nominal contenant 'raison' s'exprimerait :

```
//NP/_[lemma="raison"]
```

C'est cette fois le symbole / qui exprime la dominance, comme dans les structures hiérarchiques des chemins de fichiers.

Quoiqu'il en soit, le panorama des outils disponibles était amplement suffisant pour ce type de données : Lai et Bird (2004) présentent par exemple pas moins de six outils et langages différents pour les banques d'arbres syntaxiques.

3.3.3 Interrogation de corpus annotés par Syntex : un outil pour les linguistes ?

La question principale restant en suspens était la conversion des données fournies par Syntex à un format compatible avec les outils précédemment décrits. Il fut donc nécessaire de construire un de ces innombrables convertisseurs de format dont les pratiquants du TAL et de la linguistique outillée sont devenus (par obligation) des experts.

Le choix de l'outil spécifique que nous allions utiliser s'est porté sur TigerSearch, développé dans la tradition de CQP par l'IMS de Stuttgart (König *et al.*, 2003) [6]. Cet outil avait sur ses concurrents plusieurs avantages : il proposait plusieurs types de format en entrée (dont un format spécifique en XML que nous avons choisi), il disposait d'une interface graphique moderne, permettant de visualiser les annotations mais aussi de construire des requêtes, et surtout il utilisait une indexation robuste permettant une montée en volume (ce que ne proposaient par exemple pas les outils Tgrep ou dérivés d'Xpath).

Le convertisseur que j'ai conçu prenait donc en entrée les sorties natives de Syntex et effectuait une série de calculs pour identifier les constituants implicitement exprimés par les relations de dépendance. L'algorithme était grossièrement le suivant :

Pour chaque mot M ayant la catégorie Cat
  Si celui-ci a comme dépendants les n mots M_i avec la relation R_i (i entre 1 et n)
    Construire le sous-arbre associé à M :
      créer un nouveau nœud de catégorie "$S.Cat$"
      ajouter une branche de ce nœud étiquetée *Tête* vers le mot M
      Pour i de 1 à n :
        une branche étiquetée R_i vers le mot M_i si c'est une feuille
        vers le nœud du sous-arbre associé à M_i si M_i a des dépendants

Au final on obtient un arbre dont la représentation graphique dans l'outil de visualisation TigerSearch est visible dans la figure 3.4 (dans cette version la relation *Tête* est notée *H*, et certaines relations ont été laissées vierges).

Le nœud étiqueté *VROOT* est un artefact pour satisfaire l'exigence d'une structure arborée, et représente la racine de l'arbre. Y sont rattachés tous les constituants n'ayant pas de gouverneur identifié par l'analyseur. Si c'est une situation normale pour un verbe principal (identifié comme descendant d'un nœud *S*) ou pour les signes de ponctuation, c'est également ainsi que sont représentés les éléments non rattachés, comme on le verra plus loin.

6. www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/

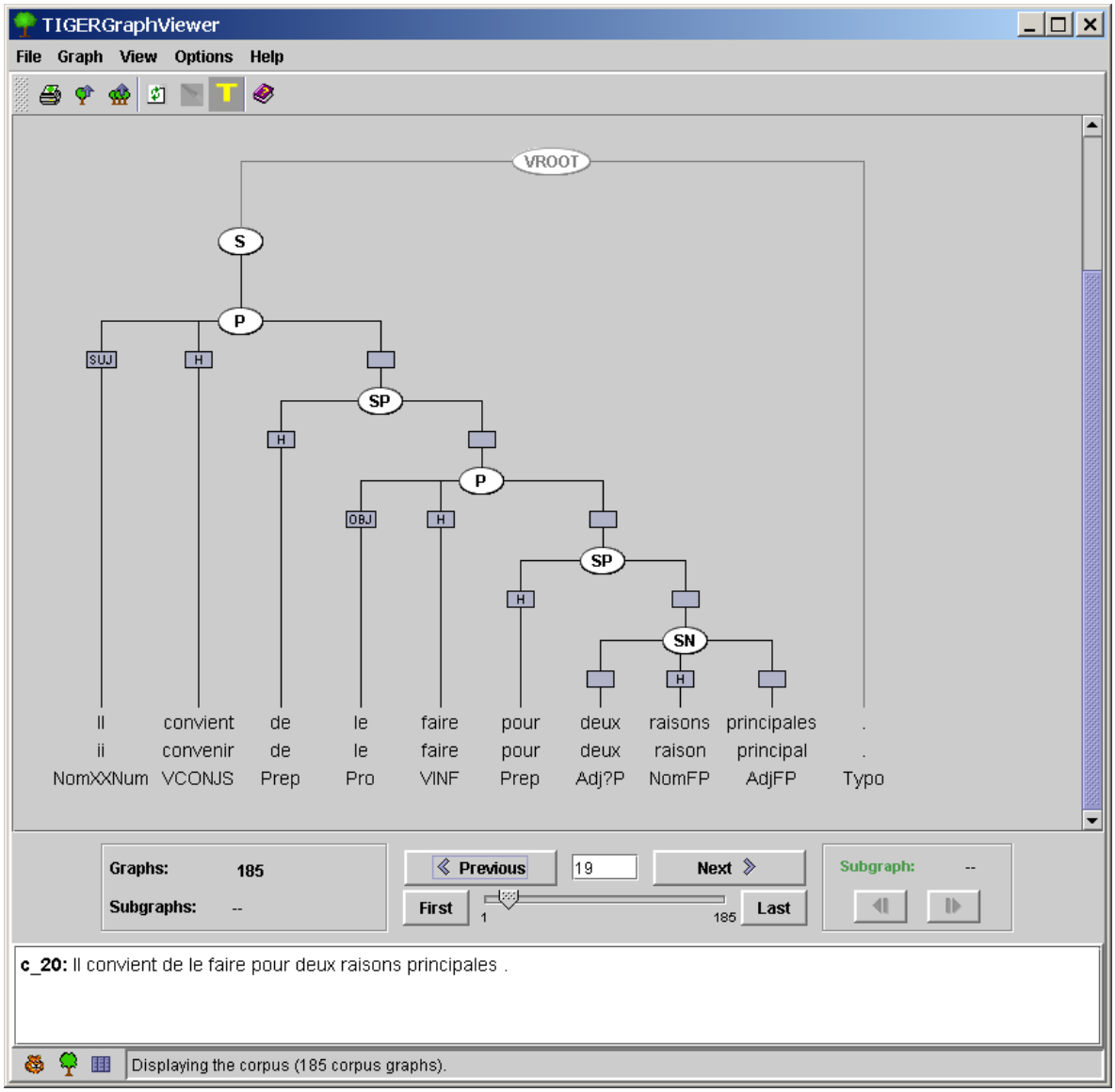

FIGURE 3.4 – Arbre syntaxique reconstitué à partir des résultats de Syntex, visualisé par Tiger Search

Une fois les corpus étiquetés par Syntex rendus compatibles avec l'outil, il était possible de les interroger en utilisant le moteur fourni avec Tiger Search. Celui-ci propose un langage de requêtes similaire à celui de Tgrep, mais avec des fonctionnalités supplémentaires. Les principales fonctionnalités et des exemples sont présentés ci-dessous :

- Chercher un mot en fonction de sa forme, son lemme et/ou sa catégorie :
  `[l=/.+able/ & c=/Adj.*/]`
  permet de trouver les adjectifs se terminant par *-able* ;
- Chercher des mots en fonction de leur position :
  `[c=/Nom.*/] . [c=/Adj.*/]`

permet de trouver un nom suivi immédiatement d'un adjectif (on peut spécifier la distance maximale en nombre de mots, ou encore rechercher toute cooccurrence ;

- Chercher une relation hiérarchique entre deux constituants ou entre un constituant et un mot : `[cat="SN"] > ["raison"]` permet de trouver les syntagmes nominaux qui contiennent (immédiatement) le mot *raison*. Là aussi on peut préciser la distance de la relation hiérarchique en nombre de constituants intermédiaires.
- Chercher deux constituants reliés par une relation précise :
  `[cat="P"] >SUJ [c=/Pro.*/]`
  permet de trouver les pronoms sujets (i.e. reliés par une relation sujet à un nœud de phrase.
- Chercher des structures à plus de deux constituants. Dans ce cas, il est nécessaire d'énumérer les contraintes une à une, en utilisant un mécanisme de mémorisation pour identifier les composants qui entrent dans plusieurs contraintes. Par exemple :
  `[cat="P"] >SUJ #A:[cat="SN"]`
  `&A > [l="raison"]`
  permet de trouver les syntagmes nominaux sujets qui contiennent le mot *raison*. Le symbole `A` est une variable (attribuée par la première partie de la requête pour être utilisée dans la seconde) permettant de préciser que le syntagme nominal sujet est aussi celui qui contient *raison*.

Une fois la requête définie, elle est projetée (de façon très rapide) sur le corpus, et l'ensemble des phrases trouvées (et les parties de ces phrases en correspondance avec la requête) sont alors visualisées comme dans la figure 3.4. De plus, des statistiques générales et une version texte des extraits correspondants sont présentés. Il est par contre difficile, comme dans tous les outils d'interrogation des corpus arborés, d'avoir facilement accès au contexte (phrases voisines).

On voit avec le dernier exemple de requête que la situation devient très vite complexe, et qu'une recherche mettant en jeu plusieurs mots pivots (comme les structures d'énoncés définitoires) nécessite un grand nombre de contraintes élémentaires. Même une simple série de 4 mots successifs va nécessiter 3 sous-requêtes de juxtaposition et deux mémorisations pour les relier.

L'interface graphique de Tiger permettant de construire les requêtes en utilisant la souris ne permettait pas vraiment de s'affranchir de ces difficultés, et nécessitait au final que l'utilisateur conçoive la requête comme un ensemble de contraintes déclarées individuellement.

Malgré les efforts fournis tant par les concepteurs d'outils comme Tiger que par les membres de l'ERSS pour les faire fonctionner et les présenter aux différents collègues de Toulouse ou de Nice, on ne peut conclure à un succès de cette proposition. Le langage d'interrogation était par trop complexe, et hormis les séances enthousiastes d'expérimentation que ne manquent pas de déclencher de nouveaux outils, à ma connaissance aucune investigation réelle de corpus n'a utilisé cet outil.

Par contre, ce travail a été très positif pour les développeurs de Syntex, qui avaient à la fois la possibilité de visualiser le résultats des analyses, d'identifier facilement les erreurs et les manques, et aussi de pouvoir présenter simplement les données produites en utilisant le module de visualisation.

3.4 Bilan et principes

Ces différents travaux de développement d'outils, mais surtout les interactions que j'ai pu avoir avec leurs utilisateurs effectifs ou potentiels me permettent de tirer quelques enseignements sur le rôle et la place de ces outils dans le travail du linguiste sur corpus.

Je commencerai par lister ce que j'estime être les compétences nécessaires pour un utilisateur de ces outils, avant d'aborder les limites de ces derniers.

3.4.1 Compétences nécessaires pour interroger des corpus

On a vu qu'il existe un impressionnant inventaire d'outils d'interrogation de corpus, et à l'heure où j'écris ces lignes je suppose que d'autres sont en train d'être conçus. Il est normal que de tels outils apparaissent et disparaissent ; comme tout programme informatique ils ont vocation à répondre à une demande qui ne peut être que temporaire : les corpus sont de plus en plus volumineux, les annotations sont de plus en plus variées et complexes (et ont une fâcheuse tendance à se cumuler), les usages et les pratiques générales de l'outil informatique évoluent et avec elles les attentes des utilisateurs, et surtout les besoins des linguistes ne cessent ne se diversifier.

La plupart de ces outils sont conçus pour être utilisés par un linguiste prototypique dont on suppose qu'il a bien compris l'intérêt de travailler sur des données annotées pour lui faire gagner du temps dans l'accès aux données. Les efforts consentis par les développeurs en termes d'interface homme-machine sont des preuves qu'ils ne s'adressent pas à des informaticiens, dont la culture (au moins pour ceux de ma génération et des précédentes) va plus vers des outils modulables et intégrables, avec des interfaces en ligne de commande. A ce linguiste on propose donc logiquement de se concentrer sur sa tâche : la définition d'un besoin et l'analyse des résultats retournés par l'outil.

On fait généralement en sorte qu'il ne soit pas confronté directement aux détails sordides de l'implémentation et des données annotées (les formats divers et variés, les truchements de stockage et de structuration des données, etc.).

Il est par contre supposé (plus ou moins implicitement) qu'il dispose d'un ensemble de compétences techniques, concernant à la fois les modes d'accès classiques aux données textuelles (comme les expressions régulières, omniprésentes dans les outils d'interrogation) et les choix qui ont été faits lors de l'étiquetage automatique.

Si le premier point rejoint des préoccupations pédagogiques dans la formation des étudiants en sciences du langage, qui fait que les expressions régulières sont une compétence absolument indispensable pour ceux-ci dès les années de licence (voir en conclusion, section 1, page 221), le second montre qu'il est important de préciser les principales caractéristiques des processus utilisés en amont pour qu'un tel outil soit utilisé au mieux de ses capacités.

Par exemple, il est important pour un utilisateur de bien connaître les règles de segmentation qui ont des conséquences sur toute la chaîne d'analyse. Savoir si une locution prépositionnelle (comme *à partir de*) est à rechercher telle quelle (à saisir comme une seule unité dans la requête) ou à décomposer en trois éléments est absolument vital quel que soit le mode d'interrogation.

De même, la plupart des patrons (morpho-) syntaxiques doivent prendre en compte les éventuelles erreurs de l'analyse. Si certaines sont inévitables, et sont simplement le prix à payer pour un accès simplifié aux données, certaines peuvent entraîner des choix lors de la définition des requêtes. Par exemple, il est préférable dans mon expérience de rechercher les formes

participiales en utilisant une expression régulière s'appliquant aux formes de surface (comme `/aimée?s?/`) plutôt que de faire appel à la catégorie morpho-syntaxique ou au lemme : les deux peuvent varier sans préavis entre la forme verbale (lemme *aimer*) ou adjectivale (lemme *aimé*). Les ambiguïtés de ce type sont légion, et croissent en fréquence avec la sophistication du traitement : il est par exemple préférable d'utiliser quand même des critères de juxtaposition dans un corpus étiqueté syntaxiquement plutôt que de recourir systématiquement aux relations syntaxiques, quand la variété des positions attendues ne le justifie pas (par exemple les déterminants).

Enfin, comme on l'a vu pour les patrons des énoncés définitoires, il est souvent nécessaire d'ajouter des contraintes parfois *ad hoc* pour affiner ceux-ci, surtout lorsque leur utilisation doit faire par la suite l'objet d'un traitement automatisé. Notamment, interdire certains mots ou classes de mots aux frontières d'une structure pour la désambiguïser est un passage obligé pour accroître la précision du repérage ; il est clair que cela n'entre cependant pas dans la définition du phénomène linguistique correspondant, mais relève bien des nécessaires ajustements de l'ingénierie linguistique.

Un dernier point concernant les compétences et capacités des utilisateurs est leur rapport à la qualité et à la quantité de résultats qu'un tel outil produit. Si l'accès à la quantité est généralement l'intérêt premier de faire appel à un traitement automatique, elle peut également être une source de découragement. Mais à l'inverse j'ai été plus souvent surpris par l'énergie déployée par des collègues pour analyser (en fait trier) manuellement des quantités impressionnantes de résultats qui m'auraient personnellement conduit à repenser ma définition de la cible pour affiner la recherche. Ce point est directement lié à la question des données bruitées : admettre leur présence inévitable dans tout résultat d'une recherche automatisée est un pas décisif dans le développement des compétences du linguiste sur corpus. Si là aussi la quantité de bruit acceptable dépend des individus, il est surtout important lorsque l'on accompagne ce genre de travaux de bien sensibiliser les utilisateurs à la distinction entre le bruit dû à une mauvaise définition du problème et celui dû aux imperfections du système (d'étiquetage notamment) et à la « propreté » des données.

Pour toutes ces raisons, je considère qu'il n'est pas vraiment possible de s'abstraire d'un lien direct entre l'utilisateur et les données brutes sorties de l'analyseur. De fait, certains outils les plus complexes prennent le soin de présenter les détails de l'analyse (afficher les tags directement dans l'interface par exemple, comme c'est le cas dans Tiger Search).

Pour finir sur les compétences impliquées dans ce type de pratique, je tiens à rappeler qu'à mon avis ce type de travail est aussi un enrichissement dans la description d'un objet d'étude, notamment par la rigueur qu'il nécessite dans la délimitation d'un phénomène.

3.4.2 Limites intrinsèques des outils

Un outil n'étant qu'un outil, il possède quelle que soit sa sophistication un ensemble de limitations que l'on doit accepter et prendre en compte lors de son utilisation. Les grandes difficultés (voire les échecs) rencontrées lorsque l'on cherche à faire utiliser un outil à un utilisateur permettent de préciser les grands types de problèmes rencontrés.

Hormis les difficultés techniques d'utilisation, dues notamment à la complexité des formalismes et à l'opacité des traitements, une partie du problème vient d'une demande souvent trop exigeante de la part des chercheurs.

Dans le cas des recherches sur corpus, j'ai bien souvent été confronté à des collègues qui s'intéressaient à des phénomènes inaccessibles par un simple traitement morpho-syntaxique.

Le cas le plus fréquent était l'appel à des considérations sémantiques dans la définition des structures visées. Que ce soit des noms *concrets* ou des expressions *spatiales*, les contraintes exprimées pour la recherche d'une structure particulière nécessitaient au minimum l'établissement de listes parfois très complexes, et étaient parfois simplement inaccessibles par des moyens automatiques. S'il est toujours possible d'approcher ces questions par des patrons très génériques, le bruit produit par la recherche était généralement à l'origine d'un découragement tout à fait compréhensible. Quant à l'élaboration de ressources intermédiaires comme les classes lexicales, l'effort nécessaire n'était que très rarement justifié par l'objectif poursuivi. Mais parfois certaines ressources directement utilisables faisaient ainsi leur apparition, certains collègues ayant dans leurs tiroirs des listes exhaustives de classes lexicales qu'il était très simple d'intégrer dans des patrons (moyennant une fonctionnalité supplémentaire des outils, comme l'intégration de listes de lemmes prédéfinies).

Une autre source d'échec du même genre, mais plus difficile à accepter, provenait des limites des analyseurs, et notamment des analyseurs syntaxiques. Ce fut le cas à plusieurs reprises, et notamment lors de la collaboration avec les collègues de Nice qui travaillaient pour certains sur des phénomènes syntaxiques comme les subordonnées en initiale de phrase. Or, ce type de structures était justement un de ceux qui étaient laissés non-analysés par Syntex (le rattachement de la conjonction ne pouvant se faire, comme le montre l'exemple de la figure 3.5 où *alors que* n'est rattaché qu'à la racine virtuelle *VROOT*).

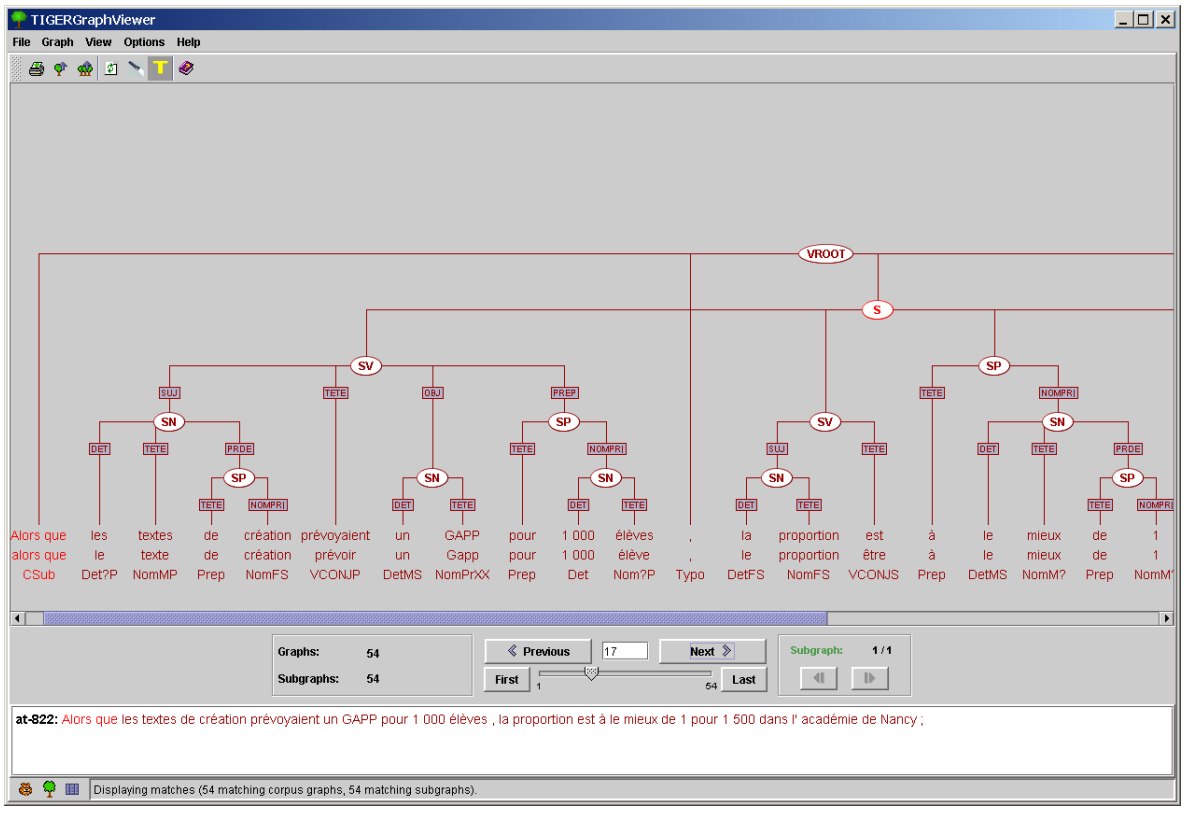

FIGURE 3.5 – Analyse partielle par Syntex

Un autre problème rencontré lors de la diffusion de TigerSearch sur les corpus syntexisés

fut d'ordre plus culturel, mais je tenais à l'évoquer ici : les arbres syntaxiques produits par le convertisseur ne correspondaient à aucun formalisme syntaxique standard, et les choix effectués par nos soins pour répondre aux contraintes de l'outil ont pu parfois être critiqués par des spécialistes de la question. Bien que cette incompatibilité ne soit pas au final bloquante, elle avait notamment comme conséquence d'éloigner comme utilisateurs les linguistes habitués à des représentations formelles des grammaires génératives (pouvant dans dans certains cas s'apparenter fortement à des requêtes). Mais dans ce type de cas, il existe des outils dédiés à la manipulation de corpus et de grammaires relevant du formalisme considéré (LFG et HPSG notamment).

Au-delà de ces limitations, dans certains cas les demandes dépassaient simplement les capacités de l'outil tel qu'il était conçu. Je citerai en exemple le travail d'Anne Condamines sur l'exploration des anaphores infidèles comme marqueur d'hyperonymie (Condamines, 2005a). Ce type de marqueurs possède des contraintes contextuelles assez complexes : un patron particulier est par exemple constitué d'un syntagme nominal démonstratif, mais dont le nom tête ne doit pas apparaître dans un voisinage gauche assez grand (disons le paragraphe courant et le précédent) pour limiter grandement le bruit. Par exemple, ce type d'occurrences était visé :

mais nous ne traiterons que de la Lune et de Mars. Si la géomorphologie s'intéresse à ***ces astres****, c'est qu'elle s'attache à tout milieu qui est de la terre.*

Ici, on peut voir que le syntagme permet de repérer une relation d'hyperonymie entre *Lune* et *Mars* d'une part et *astre* d'autre part. Par contre, s'il y a répétition du nom par un démonstratif, le phénomène n'est absolument plus présent, comme dans l'exemple ci-dessous où il s'agit d'une anaphore classique :

Le solvant a ensuite été évaporé sous vide, ce qui a laissé un résidu, en l'occurrence ***le composé*** *A.* ***Ce composé*** *A s'avéra neutre...*

Si ce type de recherche ne pose pas dans l'absolu de problèmes particuliers, il est hors de portée des outils d'interrogation de corpus évoqués dans ce chapitre (et d'autres du même type). Un des problèmes provient notamment de la gestion de la négation dans les langages réguliers, d'autant plus qu'ici la négation porte sur un ensemble de positions, et du fait que le terme proscrit est situé à droite de la négation. Pour répondre à ce besoin particulier, j'ai donc dû développer un (petit) programme ad hoc, sans envisager pour autant d'ajouter la fonctionnalité correspondante à Yakwa.

Ce cas particulier n'en est qu'un parmi de nombreux autres : il est clair qu'il n'existera jamais d'outils d'interrogation de corpus capables de répondre à tous les besoins. En plus des structures très complexes recherchées ou de contraintes particulières comme la précédente, on peut évoquer des types de données qui nécessitent un traitement particulier. C'est le cas notamment des textes structurés (que ce soit la structure logique du document, ou bien des annotations supplémentaires), des textes parallèles alignés, des dialogues, etc. Dans chacun de ces cas, des besoins nouveaux s'expriment par le fait que les informations supplémentaires ont vocation à être exploitées par une interrogation : on peut avoir envie de rechercher telles séquences de mots dans un titre ou en début de section, dans les productions d'un locuteur particulier, ou des paires de mots en traduction l'une de l'autre, etc. A chaque fois, des contraintes techniques supplémentaires viendront s'ajouter aux fonctionnalités souhaitées, et le risque est grand de vouloir produire un outil générique pour un besoin particulier.

C'est notamment ce que plaidaient Ide et Brew (2000), bien qu'en traitant plus spécifiquement du développement des corpus que des outils d'interrogation :

> « *We have argued that it falls to corpus designers to engage in the near impossible task of second-guessing the needs of future corpus users. Doing this will decrease the likelihood that we design corpora which are too narrowly focussed on the needs of particular research communities. It is conceivable that our guesses about future needs turn out to be correct, but it would be an error to tailor the design of our corpora for these needs. Rather, what is needed is an open architecture approach, which will allow future users to access corpora in the ways they find appropriate.* »
> (Ide et Brew, 2000)

L'illusion de trouver un outil adapté à un nouveau besoin, et la nécessité de se débrouiller autrement est très bien résumée par Sarah Leroy (Leroy, 2004) en conclusion de son travail spécifique sur le repérage des antonomases :

> « *Reste un fossé entre les deux approches, là où se situent les principaux besoins. Si les limites de ressources comme Frantext sont atteintes (que ce soit parce que le système de recherche ne permet pas de formuler des requêtes intéressantes, parce qu'on souhaite sortir du corpus clos, ou pour toute autre raison) et que la « confection » d'un système maison n'est pas envisageable, il est peu probable qu'existe un logiciel répondant précisément à un besoin linguistique donné. La solution qui s'offre alors au linguiste est un outillage « léger », l'obtention d'une compétence permettant d'utiliser telle ou telle fonctionnalité selon les cas et de détourner les outils dont il dispose et qu'il maîtrise. Cette démarche, qui est dans la plupart des cas personnelle, spontanée et autodidacte, si elle vient à rencontrer une démarche inverse tendant à rendre des outils et des méthodes informatiques plus lisibles au commun des linguistes, pourra peut-être infléchir certaines des pratiques de l'analyse linguistique.* »

Ce point particulier rejoint des considérations pédagogiques que j'expliciterai par la suite.

3.4.3 Suite au prochain corpus

Yakwa est à l'abandon depuis plusieurs années maintenant, cela ne veut bien entendu pas dire que la pratique des corpus annotés est passée de mode, bien au contraire. Du point de vue des outils toutefois, il me semble que l'évolution est allée vers un regroupement des besoins aux deux extrémités de l'échelle que j'ai présentée en section 3.1.

D'un côté, les études sur des corpus ciblés (de taille généralement limitée) se font de façon plus légère avec des outils comme les concordanciers, qui ont eux beaucoup progressé dans leur ergonomie et leur diffusion (par exemple, AntConc de Laurence Anthony [7] est très utilisé). Pour ce qui est d'aborder des grandes quantités de données textuelles pour y rechercher des attestations, c'est maintenant le Web qui sert de plus en plus de réservoir et les moteurs de recherche génériques d'outils d'accès, comme on va le voir plus en détails dans le prochain chapitre.

De l'autre côté, le développement de patrons complexes se fait généralement directement dans le cadre d'applications de TAL dont je parlerai par la suite, et en utilisant des plate-formes comme GATE (Cunningham *et al.*, 2011) ou LinguaStream (Widlöcher et Bilhaut, 2005) qui permettent le développement de patrons complexes sur la base d'un étiquetage morphosyntaxique. Ces plate-formes ont l'avantage de proposer un environnement complet

7. `http://www.antlab.sci.waseda.ac.jp/`

qui va au-delà du seul repérage d'énoncés spécifiques, et sont notamment conçues pour être articulées avec des traitements plus complexes comme l'apprentissage automatique ou la fouille de données, comme ce fut le cas par exemple dans la thèse de Marion Laignelet (Laignelet, 2009).

La difficulté à exploiter les corpus syntaxiquement annotés existe toujours, et ces plateformes génériques ne permettent notamment toujours pas de le faire aisément. Pour ce faire, le développement de programmes spécifiques adhoc est inévitable, mais c'est une compétence qui fait maintenant de plus en plus partie du bagage des étudiants ayant une formation en TAL, qui peuvent ainsi répondre directement aux besoins qu'ils identifient. Je reviens sur ces aspect en conclusion (section 1, page 221).

Chapitre 4

La ruée linguistique vers le Web

L'arrivée du Web dans la linguistique (en attendant l'arrivée de la linguistique dans les modes d'accès au Web) a changé tout un ensemble d'habitudes de recours aux données langagières. Comme le disaient avec un enthousiasme qui fait plaisir à voir Adam Kilgarriff et Gregory Grefenstette (Kilgarriff et Grefenstette, 2003) : « *The corpus of the new millenium is the Web.* ». Il est devenu un nouvel objet d'étude en tant que source de nouveaux types de textes, un réservoir d'attestations et une manne pour les approches quantitatives du TAL. Chacun y trouve des avantages sur les corpus classiques, que ce soit la masse, la variété des genres et des langues, l'évolution permanente, qui semblent globalement compenser les inconvénients qui lui valent de nombreux détracteurs (l'opacité du contenu et son hétérogénéité, les biais de ses modes d'accès, l'absence de représentativité, etc.).

Ma propre pratique du Web comme corpus remonte à plus de 10 ans, et a concerné principalement l'acquisition de ressources lexicales pour la morphologie extensive, qui a été l'occasion d'une collaboration avec les collègues morphologues de l'ERSS (Marc Plénat, Nabil Hathout, Michel Roché, Gilles Boyé) et d'ailleurs (Fiammetta Namer et Stéphanie Lignon de l'ATILF, Georgette Dal de STL). Bien qu'il ne s'agisse pas d'une utilisation *main stream* au sein de la communauté du « Web comme corpus », cela m'a permis de me confronter à l'ensemble de ses problématiques, et d'en suivre activement l'évolution.

Je commence ce chapitre par un panorama des usages du Web en linguistique de corpus et en TAL, puis je détaille les principaux aspects méthodologiques de l'accès à cette ressource. Je présente ensuite de façon synthétique les travaux d'acquisition lexicale avant de proposer quelques pistes pour continuer à explorer ce matériau.

4.1 Une courte histoire du Web : évolution des modes d'accès et des pratiques

Si le Web est un objet encore très jeune (il vient juste de fêter ses vingt ans), son histoire est déjà très remplie. Son apparition et son développement ont bouleversé un grand nombre d'activités humaines, et les tartes à la crème ne manquent pas pour décrire cet état de fait, des autoroutes de l'information des années 1990 aux réseaux sociaux omniprésents de la fin des années 2000. Je vais tenter ici de retracer l'impact qu'il a eu sur une grande partie des travaux en linguistique de corpus et en TAL, et de montrer comment ceux-ci ont dû également évoluer rapidement pour s'adapter à un objet sujet à de très nombreux changements.

4.1.1 Vue chronologique des usages du Web en linguistique

Je vais commencer ce tour d'horizon par une nouvelle frise chronologique qui servira de guide à ce chapitre. J'ai donc tenté de retracer dans la figure 4.1 les principaux moments de son histoire, à la fois en tant qu'objet indépendant, mais aussi dans les activités de recherche en linguistique et en TAL

Cette frise met surtout le point sur les aspects techniques et institutionnels de cette évolution, en marquant certains événements qui concernent l'évolution du Web lui-même et l'impact que ceux-ci ont eu sur la communauté scientifique.

- Sur le plan de l'**évolution générale du Web**, la frise commence bien entendu par la création du Web en 1991, et à partir de là son évolution rapide : le nombre approximatif de sites Web recensés a rapidement explosé comme l'indique la courbe. Les indications viennent des enquêtes menées par la société *Netcraft*[1] ; le nombre de pages lui-même est impossible à estimer sérieusement. Les premières années du Web ont vu se développer un grand nombre de moteurs de recherche, dont seuls quelques-uns sont désormais actifs et utilisés.
- En termes de **pratiques pour les travaux en linguistique**, je n'ai pas développé la diffusion d'outils et de ressources accessibles via le Web (à part la base Frantext, qui a rapidement su profiter de ce mode d'accès pour être utilisé par de très nombreux chercheurs). Par contre, j'ai positionné sur la frise plusieurs outils permettant une exploitation linguistique du Web : ces outils détaillés dans la suite de ce chapitre ont commencé à apparaître dès 1998 et se sont rapidement multiplié. Les questionnements sur l'usage du Web en linguistique et en TAL ont très vite été organisés par la communauté scientifique, à travers des numéros spéciaux de revues et la création de groupes d'intérêt et de conférences dédiées.
- J'insiste également dans la frise sur les **aspects techniques** de l'utilisation du Web, dont la rapide évolution a eu des conséquences importantes sur les activités scientifiques. La masse de données langagières disponible sur le Web est essentiellement accessible *via* les moteurs de recherche, si bien que ce sont leurs décisions techniques qui conditionnent les exploitations déployées en TAL et en linguistique outillée. On peut distinguer trois grandes phases : l'utilisation « sauvage » de ces moteurs par des programmes spécifiques, suivie d'une période de compromis entre l'enthousiasme des chercheurs et le coût (pour les moteurs) des interrogations massives, et enfin une période actuelle de retranchement des moteurs de recherche, qui privilégient dans le meilleur des cas la mise à disposition de la communauté scientifique de ressources statiques ou de produits dérivés, comme des bases de données de séquences de mots (n-grammes).

Si les premiers utilisateurs du Web comme source de données langagières sont les chercheurs en TAL, la communauté plus large des linguistes s'est rapidement intéressée à la question, et a pu bénéficier des avancées techniques produites par les premiers. Je commencerai néanmoins par le volet linguistique.

4.1.2 Usages du Web en linguistique : des attestations faciles d'accès aux doutes sur leur valeur

Comme je l'ai évoqué au chapitre précédent, il existe une gamme d'outils très simples, très pratiques et très faciles à installer, permettant d'avoir accès rapidement et sans apprendre de

1. `http://www.netcraft.com/survey/`

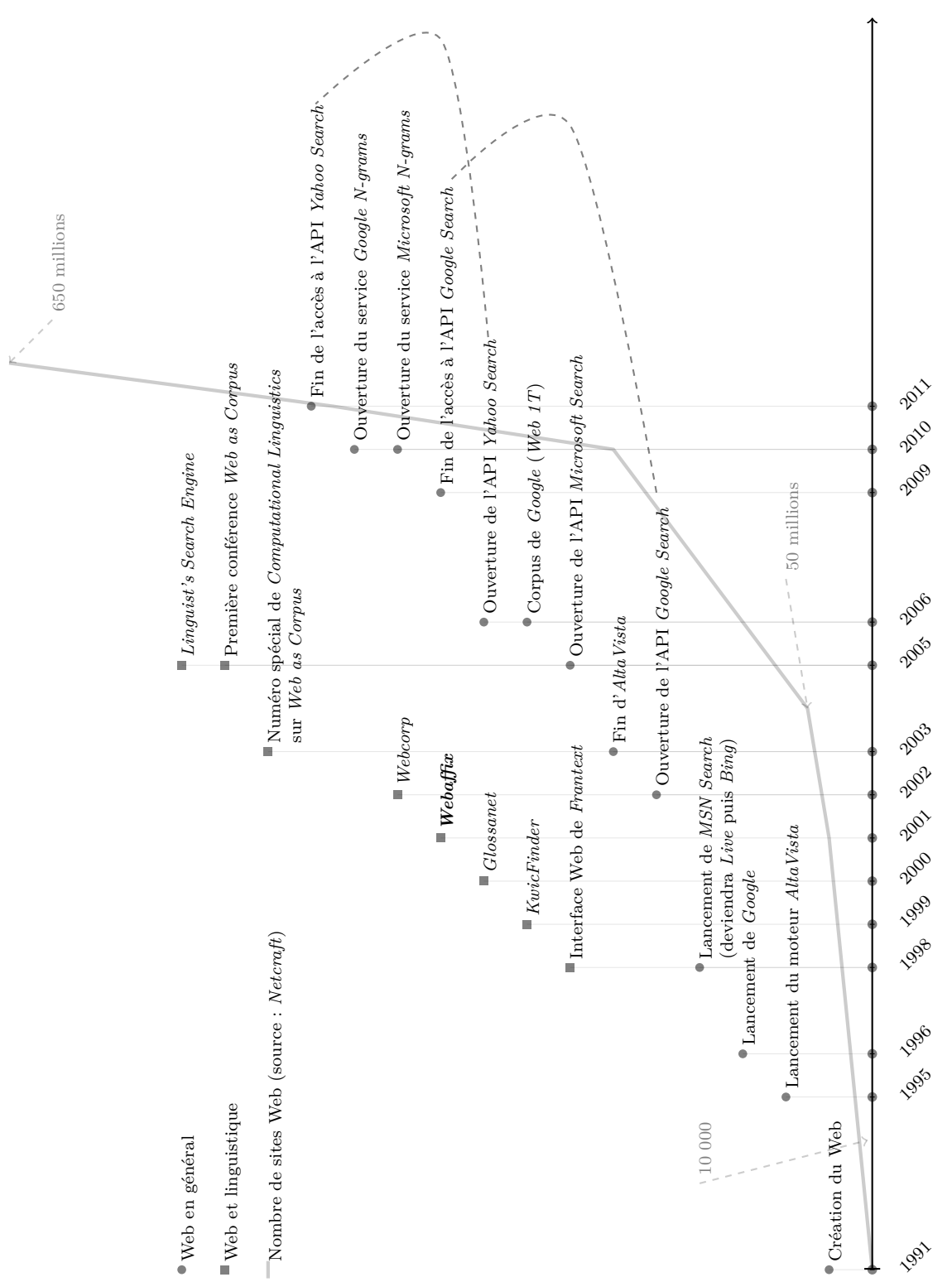

FIGURE 4.1 – Frise chronologique de l'utilisation du Web

langage de requête abscons à une quantité très importante de textes au format numérique : les moteurs de recherche sur le Web. Il n'est désormais plus rare de trouver dans des exempliers de conférence en syntaxe ou en sémantique lexicale des exemples d'énoncés dont il est dit prosaïquement qu'ils « viennent de Google », comme ces extraits de plusieurs articles de linguistique récents qui adressent des problématiques variées :

– La fragmentation syntaxique :

> *(34) Dans ce livre, il parle de deux papas et de leur enfant. Enfant, qui est victime de jugements injurieux de la part de ses camarades d'école [Google]*

– Les euphémismes :

> *(1) I don't know, I must be a dolt because I can't seem to change the color. (Google)*

– Les emplois de prépositions particulières :

> *(18b) Le village est coiffé par des champs dominant de près de 300 mètres. [Google]*

On remarquera que l'origine exacte des exemples n'est pas évoquée, ni la nature des documents dont ils sont extraits, que ce soit comme ici une critique de livre, une discussion de forum ou un journal de voyage.

Toutefois, la méfiance vis-à-vis de la source de données persiste, et bien souvent à juste titre, pour preuve cet extrait :

> « *Quand on consulte un moteur de recherche comme Google, on observe que les deux clitiques sont en effet attestés dans des cas contraires à la norme, ce qui suggère que la confusion semble concerner non seulement des stades plus anciens du français, mais la langue actuelle. Il est évident que l'utilisation de Google comme corpus demande des précautions, dans la mesure où on ne connaît pas l'identité des utilisateurs, qui peuvent en outre laisser des fautes de frappe. Une recherche sur Frantext sur une période récente et sur un corpus oral soigneusement transcrit serait évidemment nécessaire afin de vérifier l'hypothèse [...].* »
>
> (Lamiroy et Charolles, 2010)

De fait, bien souvent ce type d'exemples concerne des emplois marginaux des structures étudiées, et ils sont utilisés pour démontrer la relativité des contraintes qui sont supposées les régir, grâce à des locuteurs qui s'expriment dans un média bien moins rigide et contrôlé que les articles de journaux ou les textes littéraires qui fournissent l'essentiel des attestations. En quelque sorte, le Web comme source d'énoncés joue alors le rôle des corpus d'oral spontané difficiles à trouver et souvent très limités en volume.

Des études spécifiques sur le Web (comme celle de Resnik *et al.* (2005)) montrent même son intérêt spécifique pour trouver des exemples massifs de structures syntaxiques jusqu'ici déclarées impossibles (on verra le cas également en morphologie dans ce chapitre).

Un autre usage très classique des moteurs de recherche tel qu'il transparaît dans un grand nombre de publications de linguistique est l'utilisation des fréquences telles qu'indiquées par le nombre de pages renvoyées pour une requête. Ces valeurs sont utilisées pour indiquer que le phénomène visé est fréquent (« *plus de 10.000 occurrences selon Google* »), ou au contraire très rare (« *aucune occurrence, même sur Google* »), ou encore pour comparer deux emplois concurrents. De nombreux débats ont eu lieu sur la fiabilité des chiffres renvoyés par les moteurs de recherche. Le plus célèbre repose sur l'expérience menée par Jean Véronis [2] qui a

2. Voir à ce sujet son blog sur `http://blog.veronis.fr/`

mis au jour le manque total de cohérence de ces résultats pour les principaux moteurs de recherche. Quoiqu'il en soit ces valeurs sont, lorsqu'elles indiquent des variations très importantes, une estimation à peu près aussi fiable que celles fournies par un corpus « classique ». Pour plus d'exemples d'études de ce type, voir notamment (Lüdeling *et al.*, 2007) et (Hundt *et al.*, 2007).

Dans le même ordre d'idée, le Web peut également être utilisé pour des études comparatives, en ciblant des productions de différentes communautés linguistiques, comme le fait Wooldridge (2004) lorsqu'il compare le français de France et celui du Québec, opération rendue très simple par la configuration d'un moteur de recherche ou la spécification d'un suffixe de domaine particulier (`fr` versus `ca` par exemple).

Un autre type d'étude concerne cette fois le Web comme objet, et non plus comme outil ou simple source d'exemples diversifiés. De nombreux travaux visent à observer et caractériser les nouveaux modes d'expression induits par le développement rapide des moyens de communication. Que ce soit les discussions en ligne, les forums ou les blogs, ces données facilement accessibles permettent d'étudier le comportement langagier (plus ou moins) spontané de communautés, et renouvellent certaines questions en analyse linguistique à différents niveaux (voir par exemple Mourlhon-Dallies *et al.* (2004)). On y voit notamment des études sur les écarts à la norme dans ces modes d'expression, des analyses conversationnelles, ou encore des approches sociolinguistiques s'intéressant à la formation de communautés autour de ces lieux particuliers d'échange. Tous ces travaux semblent exprimer une certaine circonspection par rapport aux objets étudiés, et une question récurrente concerne l'existence ou non d'une réelle nouveauté des pratiques langagières.

Le même type de question est soulevée par les travaux qui cherchent à inventorier et à caractériser les *genres du Web*. Je citerai notamment ceux de Marina Santini (Santini, 2007) qui propose de voir dans ces nouveaux objets textuels une hybridation de genres existants, mais laisse la question ouverte, tout en approchant également cette problématique par le biais d'une classification automatique en genres des pages Web. Cette problématique des genres se décline également à des niveaux plus locaux, comme les travaux de Valette et Rastier (2006) dans le cadre du projet Princip qui visait la caractérisation linguistique et la détection automatique de sites Web racistes.

Il est donc rassurant de constater que la communauté scientifique semble aborder ces données avec dynamisme et circonspection, mais en évitant un enthousiasme aveugle face au manque de fiabilité des données. Les autres utilisateurs du Web comme source de données, les chercheurs en TAL et autres développeurs d'applications d'ingénierie linguistique font peut-être étalage de moins de doutes, tant sont nombreuses et apparemment sans limites les utilisations qui peuvent être faites de cette masse de données.

4.1.3 Usages du Web en TAL : apologie de la quantité et de la diversité

Le Web et le TAL vivent une histoire commune depuis l'apparition du premier, ne serait-ce que par la problématique de la recherche d'information et les besoins en ingénierie linguistique que les applications ont fait émerger. Je me concentrerai ici sur une petite partie des travaux, en ciblant ceux qui (comme nous l'avons fait à l'ERSS) cherchent à extraire du Web des données linguistiques exploitables.

Les premiers chercheurs de TAL à avoir su profiter du Web semblent être ceux qui s'intéressaient à l'ingénierie multilingue, que ce soit pour l'acquisition de lexiques de transfert ou pour alimenter des systèmes de traduction automatique statistique (Grefenstette, 1998;

Resnik, 1999). Un des avantages du Web est en effet, en plus de la variété des langues qui y sont représentées (malgré l'hégémonie évidente de l'anglais), le grand nombre de sites traduits en plusieurs langues (qu'ils soient institutionnels ou commerciaux). Des indices explicites et fiables peuvent être utilisés pour repérer qu'une page Web contient la traduction d'une autre, et permettent de construire automatiquement un corpus parallèle.

Du côté des ressources monolingues, on peut citer la récolte automatisée et à grande échelle d'entités nommées (Jacquemin et Bush, 2000) et nos propres travaux sur l'extension de lexiques morphologiques et la découverte de néologismes (Hathout et Tanguy, 2002; Hathout et Tanguy, 2005; Hathout *et al.*, 2009).

Les exemples ne manquent pas non plus du point de vue plus *hard-core* du TAL, à savoir l'exploitation de grandes quantités d'informations de bas niveau extraites de données textuelles non enrichies. La plupart des besoins pour l'entraînement de systèmes statistiques de TAL se résument à des séquences courtes de mots (n-grammes) avec une fréquence associée. Les utilisations de ces données très frustes (appelées *modèles de langue*) recouvrent le traitement de bas niveau comme l'étiquetage morpho-syntaxique, l'analyse syntaxique, la désambiguïsation, mais aussi les calculs plus sémantiques sur des principes distributionnels (classification d'unités lexicales, repérage de relations). Le besoin principal de ces types d'approches est une quantité très importante de données, qui compense en partie la pauvreté des informations. Le Web est donc un choix logique pour accumuler ces grandes quantités de n-grammes. De plus, les compagnies qui construisent et gèrent les principaux moteurs de recherche Google (Brants et Franz, 2006), Microsoft (Wang *et al.*, 2010) et Yahoo ! ont récemment mis à disposition de la communauté scientifique de telles données, dont le succès a été immédiat. Voir notamment (Lin *et al.*, 2010) pour des exemples d'utilisation de ces données (ici de Google) et Keller et Lapata (2003) pour l'exploitation des n-grammes avant la distribution des données des moteurs.

Des approches plus fines peuvent concerner également l'interrogation du Web pour désambiguïser le rattachement prépositionnel en analyse syntaxique (Gala, 2003), pour vérifier la validité d'une dérivation morphologique (Namer, 2003) ou encore pour trouver des reformulations (Duclaye *et al.*, 2002).

Dans tous ces travaux, le Web est vu comme un réservoir indifférencié de textes à analyser, indépendamment des grandes questions sur leur nature. Cette caractéristique des approches ultra-massives des données en TAL se retrouve également dans le mouvement actuel en linguistique de corpus visant la constitution de corpus génériques de plus en plus volumineux : on a vu au chapitre précédent à travers la courbe de la frise chronologique de la figure 3.1 (page 54) que les corpus les plus volumineux sont issus du Web.

La remise en cause de la nature et de la qualité des données y est nettement moins présente que pour les usages directs par des linguistes (qui, eux, regardent bien évidemment les données de plus près). Quelques exceptions sont à noter toutefois, notamment lorsque certaines expériences mettent au jour des résultats aberrants. Les données fournies par Google notamment, ont causé quelques émois dans la communauté, lorsque Hal Daumé [3] a découvert que les séquences de cinq mots les plus fréquentes correspondants au schéma *the X Ved the Y* étaient dans l'ordre (et avec des fréquences de plusieurs dizaines de milliers d'occurrences) :

the surveyor observed the use
the rivals shattered the farm
the link entitled the names

3. `http://nlpers.blogspot.com/2010/02/google-5gram-corpus-has-unreasonable.html`

the trolls ambushed the dwarfs
the dwarfs ambushed the trolls

Il semblerait que ces fréquences disproportionnées pour des séquences aussi improbables soient le fait des nombreuses pages Web de *spam* générées par milliers pour leurrer les moteurs de recherche et amener leurs utilisateurs vers des sites souvent peu recommandables. Comme on le verra plus loin, il s'agit d'un des nombreux problèmes du Web, mais qui peut avoir des conséquences si les données sont utilisées sans discernement.

Pour reprendre la notion de complexification des données et des techniques, il est clair au vu des quelques exemples de pratiques décrits dans cette première section que le Web est un lieu très significatif des changements que je cherche à dégager dans ce mémoire. Du point de vue de la linguistique empirique, les données accessibles via le Web sont certes nombreuses, mais surtout très difficiles à décrire et à exploiter, et posent un ensemble de questions plus épistémologiques sur le statut des données elles-mêmes. Le TAL a permis de développer un ensemble assez vaste (étant donnée la brièveté de la période) d'outils pour permettre leur exploitation : si ces outils ne sont pas eux-mêmes d'une grande complexité de conception ni d'utilisation, ils ont amplifié les problèmes liés à la nature et au contrôle des données manipulées, les rendant elles aussi encore plus difficiles à évaluer, et produisant des résultats qui posent un ensemble de problèmes dans leur validation.

4.2 *Web for corpus* versus *Web as corpus*

Avant de détailler les aspects plus techniques qui sous-tendent les travaux qui exploitent le Web, il convient d'en faire une première typologie, en plus de la distinction des objectifs entre linguistique et TAL. De Schryver (2002) propose de distinguer les approches du Web pour constituer un corpus (Web *for* corpus) de celles qui considèrent l'ensemble du Web accessible comme un seul gros corpus (Web *as* corpus). Ces deux façons d'aborder la question ont des implications méthodologiques importantes, même si dans certains cas on peut remplacer l'une par l'autre. Je commencerai par quelques précisions terminologiques sur la notion de corpus dans ces deux locutions.

4.2.1 Abus du terme « corpus »

De nombreux débats ont abordé la définition de ce qu'est un corpus en linguistique, et se sont tout naturellement adressés au cas particulier du Web. La tendance généralement exprimée est de refuser le statut de corpus au Web, pour plusieurs raisons. Sinclair (2004) est sans doute le plus radical (dans les citations qui suivent, c'est moi qui souligne) :

> « *The World Wide Web is* not a corpus*, because its dimensions are unknown and constantly changing, and because it has not been designed from a linguistic perspective.* »

Rundell (2000) est un peu moins sévère :

> « *The big question here is about the actual value of the web as a corpus. In fact, of course, it is* not a corpus at all *according to any of the standard definitions : what it is is* a huge ragbag of digital text*, whose content and balance are largely unknown. [...] So the first caveat is that the web should not be regarded as a representative*

> *sample of English (or any of its other languages), and cannot therefore be used as a basis for making reliable generalizations about linguistic behaviour.* »

Ces deux remarques sont à mettre en regard de la tradition anglaise de la constitution de corpus de référence (le BNC en étant l'étalon), construit comme représentatif d'un état de langue en veillant à l'équilibre quantitatif des différents genres de texte, en plus de la description précise de l'origine des textes.

Dans une autre tradition, mais tout aussi radical, Rastier (2005) met en avant la notion plus générale de critère linguistique pour pouvoir attribuer le statut de corpus à une collection de textes :

> « *De fait, tout regroupement de textes ne mérite pas le nom de corpus. Ainsi une banque textuelle peut regrouper des textes numériques de statuts divers : aucun critère linguistique ne permet cependant leur totalisation, sauf l'hypothèse que la langue leur confèrerait une unité a priori ; mais même organisée en base de données, une banque textuelle ne devient pas pour autant un corpus.*
>
> *Un hypertexte n'est pas non plus par nature un corpus : soit c'est l'équivalent numérique d'un codex dont les renvois internes sont des liens hypertexte ; soit c'est l'hypertexte indéfini et il se confond avec le web – qui n'est* pas un corpus *mais, un euphémisme s'impose,* une aire de stockage, voire une décharge publique*.* »

Il faut donc considérer le Web comme, au mieux, une simple collection de textes dont la nature, la taille et la distribution sont mal connues (voire inconnaissables). Par contre, dans les usages concrets qui en sont faits dans les travaux de linguistique descriptive et de TAL, il semblerait qu'il soit utilisé en tant que corpus, au sens d'un ensemble de productions langagières exploitées à des fins linguistiques.

L'abus de langage est donc de mise (même si regrettable) dans la plupart des travaux décrits ici, mais avec toutefois une variation importante d'une situation précise à une autre.

4.2.2 Constituer un corpus à partir du Web

L'utilisation du Web comme *source de corpus* implique que le chercheur va utiliser une collection de textes extraits du Web. Dans ce type de cas, l'usage du terme corpus est moins choquant, puisque le linguiste a la possibilité de définir lui-même, et sur des critères qu'il est en mesure d'expliciter, la constitution de la collection de documents.

Si cette activité peut très bien se faire à petite échelle, en glanant des pages Web (ou des sites entiers), elle peut également se faire de façon plus massive en faisant appel à des processus automatisés.

Différents outils ont en effet été proposés pour assister le repérage, le nettoyage et dans certains cas l'étiquetage de pages Web. Les outils proposés sont par exemple BootCat (Baroni et Bernardini, 2004; Sharoff, 2006) qui à partir d'une liste de mots-clés génère un ensemble de requêtes automatisées et les soumet à un moteur de recherche pour télécharger des pages (en identifiant de façon itérative de nouveaux termes de requête). A ce travail de repérage s'ajoutent des procédures de normalisation des formats (extraction du seul contenu textuel des pages, harmonisation des codages de caractères, etc.), et la possibilité d'effectuer un étiquetage automatique). Le fait de pouvoir préciser un ensemble de mots-clés de départ, et la procédure de *bootstrapping* pour étendre ceux-ci permet de cibler des corpus thématiques.

De la même façon le *GrosMoteur* de Kim Gerdes [4] effectue quant à lui un parcours du Web (*crawling*) et intègre un outil d'interrogation des pages moissonnées avec une interface dédiée.

Ces efforts techniques importants sont pour la plupart dûs à la création d'un groupe d'intérêt de l'association ACL (*SIG-WAC, Special Interest Group on the Web as Corpus* [5], organisateur notamment de la conférence WAC). Le groupe Wacky [6] a également construit par ce type de méthodes des corpus génériques pour les principales langues occidentales et les a mis à la disposition de la communauté. On peut ainsi disposer gratuitement de corpus de plusieurs centaines de millions de mots (voire plus) prêts à l'emploi (Baroni *et al.*, 2009), pour peu que l'on dispose des moyens de stockage. Certains de ces corpus sont également interrogeables directement en ligne.

Dans ces cas-là, par contre, il est évident que la notion traditionnelle de corpus est encore plus malmenée : un moissonnage automatique ne permet guère de contrôler quoi que ce soit dans la sélection des textes, que celui-ci soit déclenché par le chercheur lui-même, ou bien qu'il se serve de masses de textes prêtes à l'emploi.

4.2.3 Corpus et outils prêts à l'emploi

Je me contenterai ici de signaler quelques outils qui proposent un accès direct à des corpus issus du Web.

Le plus complet de ces outils n'est malheureusement plus en activité, mais a représenté l'effort le plus abouti pour fournir un accès à des corpus issus du Web. Il s'agit du *Linguist's Search Engine* de Philip Resnik et Aaron Elkiss de l'université du Maryland [7] (Resnik *et al.*, 2005) qui proposait un accès direct avec une interface innovante à une importante collection de pages Web étiquetées et parsées.

Serge Sharoff et ses collègues de l'université de Leeds [8] proposent une série de corpus extraits du Web, étiquetés et lemmatisés, que l'on peut interroger avec l'outil CQP (voir section 3.2.2, page 61)). Ils proposent des corpus dans plusieurs langues, dont le français.

Adam Kilgarriff propose à travers son Sketch Engine [9] l'interrogation de nombreux corpus issus du Web, en mettant l'accent sur des utilisations en lexicographie, dans la tradition britannique (Atkins et Rundell, 2008). On y trouve ainsi de nombreuses fonctionnalités, en plus de la recherche par patrons, comme l'examen des collocations et une analyse distributionnelle. Malheureusement, l'accès à ce service est payant.

Glossanet est un des services proposés depuis longtemps, et sa dernière version (Fairon *et al.*, 2008) vise spécifiquement des sources de données dynamiques du Web sous la forme de flux RSS. Par un système d'abonnement (gratuit), la dernière version de Glossanet [10] permet à un utilisateur de sélectionner (ou de définir) un ensemble de flux (dépêches d'agence de presse, journaux, blogs, etc.) et d'appliquer dynamiquement à leurs contenus une requête sous la forme d'un patron morphosyntaxique. Au final, ce service permet d'effectuer une sorte de veille linguistique en exploitant spécifiquement les aspects les plus dynamiques (et également les mieux organisés) du Web.

4. `http://grosmoteur.elizia.net/`
5. `http://www.sigwac.org.uk/`
6. `http://wacky.sslmit.unibo.it/`
7. `http://lse.umiacs.umd.edu/`
8. `http://corpus.leeds.ac.uk/internet.html`
9. `http://www.sketchengine.co.uk/`
10. `http://glossa.fltr.ucl.ac.be/`

Une dernière source de données issues du Web peut également, dans certaines circonstances, être obtenue *via* un partenariat avec les opérateurs des moteurs de recherche eux-mêmes, qui sont bien entendu les mieux placés pour disposer de données, puisque leur activité consiste justement à les trouver et à les indexer. Si les acteurs majeurs comme Google et Yahoo ne proposent pas directement ces données (uniquement les n-grammes comme indiqué plus haut), nous avons notamment eu la chance de pouvoir collaborer avec la compagnie Exalead et disposer d'un échantillon important de pages Web issues de leur index.

L'autre façon d'aborder le Web (*as* corpus) est d'utiliser directement un moteur de recherche pour accéder aux données, et c'est ce dont je vais parler plus en détails dans les prochaines sections, puisque c'est le mode d'accès que j'ai privilégié pour mes travaux.

4.3 Utilisation des moteurs de recherche : un passage obligé

Les moteurs de recherche sont des outils très pratiques pour les interrogations à la volée et la recherche d'attestations, mais ils introduisent un biais dans l'accès aux données que constitue le Web dans son ensemble. Dans tous les cas, il s'agit d'un point de passage obligé de la quasi-totalité des exploitations du Web comme corpus ou source de corpus. Le TAL entretient donc des rapports compliqués avec ces outils, objets de nombreuses critiques et sources de frustrations, d'autant plus que les intérêts scientifiques se heurtent souvent à des principes industriels et commerciaux, et de plus en plus à une concurrence entre la recherche académique et les équipes de TAL dont disposent désormais les compagnies qui créent et exploitent ces moteurs (ce que le nombre important de transferts de chercheurs de la recherche publique vers ces structures traduit bien). Dans cette section un peu plus technique nous allons voir concrètement les modes d'utilisation de ces outils pour accéder au Web comme corpus

4.3.1 Utilisation directe : la *googleologie*

On l'a vu, l'utilisation directe d'un moteur de recherche pour rechercher des attestations d'une unité lexicale ou d'une structure particulière est désormais une opération très courante. Mais il est clair qu'un moteur de recherche sur le Web n'est pas un concordancier ni un système d'interrogation de corpus, et que ses fonctionnalités sont limitées pour plusieurs raisons :

- Ce sont des systèmes de recherche d'*information*, et leur objectif est d'accéder au contenu (à la *sémantique* des textes comme disent souvent les informaticiens travaillant dans ce domaine). Si bien que les fonctionnalités de ces moteurs pour faciliter cet accès sont souvent des obstacles lorsque l'on s'intéresse spécifiquement à la forme : impossibilité de prendre en compte les variations de casse ou la ponctuation, gestion automatique de la flexion, correction automatique d'erreurs, utilisation d'équivalences lexicales diverses, etc. Parfois même les termes de la requête sont purement et simplement absents des documents renvoyés, puisque d'autres mécanismes sont déployés pour associer un document à une requête (dont le fameux principe d'utilisation des termes dans les liens hypertextes qui pointent sur un document pour indexer celui-ci, au lieu de son contenu).
- Ce sont des systèmes qui doivent gérer une quantité impressionnante de données, et qui ont une exigence d'efficacité et de rapidité. Les modes d'accès complexes au matériau textuel vus au chapitre précédent sont des sources notoires de ralentissement du processus de recherche : une simple expression régulière plutôt qu'une forme exacte entraîne un

temps de calcul largement supérieur. Il existe donc au final très peu de fonctionnalités en dehors de la recherche d'une séquence de formes précises.
– L'unité de données pour un moteur est une page Web, pas une phrase ni un segment de texte quelconque. Cela signifie (en plus du problème que les fréquences affichées ne peuvent donc correspondre à des occurrences) qu'il est difficile de contrôler la portée d'une requête, ou encore d'accéder directement au contexte recherché.

Cela n'empêche pas ces outils d'être tout à fait utilisables pour des recherches portant notamment sur des unités lexicales simples, ou pour des locutions. Lorsqu'il s'agit de séquences plus complexes, s'apparentant à des patrons lexico-syntaxiques, les choses deviennent plus compliquées. Mais il reste encore quelques possibilités, notamment par le biais de requêtes utilisant des jokers, comme :

```
"plus on * plus on"
```

qui permet de rechercher des séquences dans lesquelles l'étoile peut être une série de mots quelconques (d'une taille non contrôlable, mais généralement de 1 à 5 mots).

L'absence de fonctionnalités au-delà de cette possibilité de définir des séquences *à trous* peut entraîner l'emploi de diverses ruses, et généralement la multiplication des requêtes, c'est ce qui a notamment énervé Adam Kilgarriff dans (Kilgarriff, 2007) :

> « *Working with commercial search engines makes us develop workarounds. We become experts in the syntax and constraints of Google, Yahoo, Altavista etc. We become googleologists. The argument that the commercial search engines provide low-cost access to the web fades, as we realise how much of our time is devoted to working with and against the constraints that the search engine imposes.* »

Ces remarques pertinentes entraînent Kilgarriff et d'autres avec lui à proposer de se détourner de ces moteurs pour construire des corpus à partir du Web et à développer des outils d'interrogation comme ceux que la communauté a maintenant l'habitude d'utiliser (voir quelques exemples en section 4.2.3, page 87).

4.3.2 Services intermédiaires : les concordanciers du Web

Une autre alternative aux corpus statiques construits à partir du Web pour pallier la pauvreté linguistique des moteurs de recherche est représentée par plusieurs outils en ligne qui se proposent comme intermédiaires entre le chercheur et le moteur.

C'est le cas de Webcorp [11] (Kehoe et Renouf, 2002), le premier outil de ce type qui présente sous forme de concordances les résultats d'une requête à un moteur de recherche classique (au choix de l'utilisateur). De façon assez légère et rapide, il transmet la requête de l'utilisateur, et parcourt les documents renvoyés par le moteur pour extraire les contextes d'occurrence du ou des termes choisis, ainsi que la liste des cooccurrences. Bien qu'assez rudimentaire et ne proposant pas de nettes améliorations de la syntaxe de recherche, il constitue tout de même un progrès certain par rapport à l'interrogation directe des moteurs.

Le Web Concordancer (ou KwicFinder) d'Alan Fletcher [12] fonctionne exactement sur le même principe (Fletcher, 2006).

Ces deux outils exploitent la possibilité d'automatiser les requêtes transmises à un moteur de recherche, ce qui peut permettre d'autres types d'applications.

11. http://www.webcorp.org.uk/
12. http://webascorpus.org/

4.3.3 Accès automatisé aux moteurs de recherche : les *API*

Aux premiers temps de l'utilisation du Web comme corpus, les exploitations des moteurs de recherche se faisaient au travers d'outils adhoc qui jouaient en quelque sorte le rôle du navigateur Web qu'utilise un utilisateur normal. Pour une requête donnée (respectant la syntaxe du moteur visé), le programme devait construire l'URL d'interrogation correspondante, la transmettre au serveur du moteur de recherche, récupérer la réponse (sous la forme d'une page HTML) et l'analyser pour en extraire les résultats pertinents (nombre de pages, titres, adresses et si possible extraits des documents). Comme bien d'autres à cette époque (je parle des années 1999-2002), j'avais conçu de tels programmes pour différentes applications, notamment pour l'outil Webaffix dont je parlerai plus en détail dans la prochaine section. Le développement et le maintien de tels outils était très fastidieux, notamment parce que le moindre changement impromptu de la part du moteur (dans sa syntaxe d'interrogation, mais surtout dans sa façon de présenter les résultats) nécessitait une mise à jour de tout ou partie du système, il fallait donc très régulièrement en vérifier la stabilité.

L'intérêt de ce type d'outil était bien entendu de traiter de grandes quantités de requêtes, par exemple pour calculer la fréquence (ou une estimation de celle-ci, voir plus haut) d'un terme ou d'une expression, parfois à l'échelle d'un lexique entier. Dans ce cas, le nombre de requêtes pouvait bien entendu dépasser la centaine de milliers. Certains moteurs étaient plus circonspects que d'autres par rapport à cette sollicitation massive. C'est Google qui le premier a purement et simplement interdit ce genre de pratiques en les détectant et en réagissant par une interdiction d'accès temporaire (je peux maintenant l'avouer, des années plus tard, si l'université de Toulouse 2 s'est vue privée de Google pendant quelques heures c'était de ma faute).

En contrepartie, les services de Google ont proposé un mode d'accès spécifique à leur moteur pour de telles utilisations sous la forme d'une API (un protocole informatique pour permettre la communication entre deux programmes). Ce service gratuit permettait, sur simple inscription, à la fois d'éviter les sanctions précédentes et d'accéder directement aux résultats sans avoir à fouiller dans la page de présentation destinée aux utilisateurs « normaux ». En contrepartie, le nombre d'interrogations était limité à 1000 requêtes par jour. Les autres moteurs de recherche concurrents, Yahoo ! puis Live (le moteur de Microsoft, désormais appelé Bing) lui ont emboîté le pas, avec des conditions similaires quoique généralement plus intéressantes en termes de nombres de requêtes autorisées.

Malheureusement, la tendance actuelle est à l'arrêt de ces modes d'accès, comme indiqué dans la frise de la figure 4.1. Google a rapidement suspendu ce service (d'abord en ne distribuant plus de codes d'accès, puis en le supprimant totalement), et c'est maintenant Yahoo ! qui vient de fermer ses portes. On peut supposer que le troisième et dernier va rapidement finir par les imiter [13].

Comme je l'ai évoqué plus haut, ce genre d'événements totalement indépendants de notre volonté met parfois fin à des efforts de recherche et de développement de plusieurs années. Les enjeux économiques qui entourent les moteurs de recherche sont désormais tels que la collaboration devient simplement impossible. Les fermetures de ces services mettent à mal l'ensemble des approches déclinées dans les sections précédentes, de la création d'un corpus à la volée jusqu'à l'interrogation indirecte et enrichie des moteurs.

Les alternatives sont difficiles à mettre en place : la création d'un moteur de recherche, ou plutôt d'un *crawler* capable de parcourir le Web pour en extraire le contenu textuel est un

13. À l'heure où je termine ce mémoire, l'API de Bing va en fait devenir prochainement payante.

travail de très longue haleine, et le coût matériel de son fonctionnement est colossal, bien hors de portée des budgets académiques. Il est fort possible que l'âge d'or du Web comme corpus soit derrière nous. Si c'est le cas, peut-être constaterons-nous au final que l'intense activité tant en TAL qu'en linguistique aura surtout concerné une augmentation du volume et de la variété dans les données utilisées plus qu'un changement de paradigme.

4.4 Webaffix : exploiter le Web pour une morphologie extensive

Mon expérience personnelle de l'exploitation du Web comme corpus est légèrement en dehors des principaux efforts de la communauté tels que je les ai présentés dans les sections précédentes. Hormis quelques travaux consistant à construire des corpus ciblés (notamment pour un projet consistant à étudier l'emploi de termes recommandés dans différents types de sites Web (Rebeyrolle *et al.*, 2007), ou pour différents mémoires réalisés par des étudiants de master), l'essentiel de mes efforts ont porté sur l'exploitation du Web pour y rechercher des créations lexicales. Ces travaux ont beaucoup évolué car ils ont subi de plein fouet les évolutions technico-culturo-économiques du Web et des moteurs. Dans le contexte le plus favorable (de 2001 à 2003), il m'a donc été possible (avec Nabil Hathout) de construire et de distribuer Webaffix, un outil complet et opérationnel qui a rendu de nombreux services à la (petite mais active) communauté des morphologues qui s'intéressent à la morphologie dérivationnelle et qui perçoivent l'intérêt d'utiliser des données volumineuses.

4.4.1 Objectifs et principes

Comme je l'ai présenté au chapitre 2, l'idée initiale a germé lors d'une discussion avec Marc Plénat, et nous avons voulu voir s'il était possible de repérer automatiquement de nouveaux adjectifs dérivés en *-esque* pour compléter l'impressionnante collection qu'il avait déjà réunie au fil des années dans différents corpus (Plénat, 1997). M. Plénat avait déjà vu l'opportunité d'exploiter le Web, et passait de longues heures à tester l'existence de nouveaux dérivés dont il devait auparavant supposer l'existence. L'idée de Webaffix était de remplacer ses hypothèses par une méthode purement inductive.

Il est important de noter que dès le début les phénomènes visés dans ces travaux sont très rares : si la création lexicale (et notamment par suffixation) est un phénomène relativement courant à l'échelle de l'évolution de la langue, lorsque l'on aborde des corpus les fréquences observées sont très basses. Nos dernières estimations (Hathout *et al.*, 2009) pour l'ensemble des suffixes déverbaux de noms d'actions (des suffixes très productifs comme *-tion*, *-ment* et *-age*) est que moins d'une page Web sur 200 contient une nouvelle forme lexicale de ce type.

La première version de Webaffix exploitait une fonctionnalité unique des moteurs de recherche de l'époque : la possibilité d'utiliser des troncations dans les termes de la requête. Seuls deux moteurs proposaient ce type d'accès au contenu : Altavista et Northern Light. Cette fonctionnalité a malheureusement disparu des deux moteurs au fil de leurs changements de propriétaires, de logiciel et/ou de positionnement sur le marché. Il était en effet possible de taper comme requête une chaîne comme `xxx*esque` et de voir le moteur l'interpréter comme *tout mot commençant par* `xxx` *et se terminant par* `esque`. Il était toutefois nécessaire de préciser les trois premières lettres (ou les quatre premières pour Northern Light) afin de limiter la complexité du calcul. La solution était donc simple : il suffisait de générer l'ensemble des

combinaisons de lettres envisageables à l'initiale, de décliner les requêtes correspondantes et de s'arranger pour ne pas prendre en considération les formes déjà connues (qu'elles soient des entrées d'un lexique générique ou les formes nouvelles déjà répertoriées) et le tour était joué. Ce n'était rien de bien complexe une fois qu'on avait réussi à automatiser l'interrogation du moteur, en utilisant un programme spécifique comme indiqué plus haut.

Le principe de cet outil a ensuite été étendu pour y ajouter un module d'analyse morphologique des dérivés rapportés, afin d'en identifier la base, et ainsi proposer en sortie des informations morphologiques plus complètes. C'est sur ce point qu'est intervenue la collaboration avec Nabil Hathout, puisqu'il travaillait déjà sur l'automatisation des processus d'analyse morphologique et disposait de méthodes opérationnelles et robustes pour cette tâche (Hathout, 2000).

La tâche principale nécessaire pour rendre cet outil utilisable concernait la lutte contre la quantité impressionnante de bruit généré par cette méthode.

4.4.2 Problématique du filtrage

On continue parfois à parler du Web comme *poubelle planétaire*, et certains linguistes ont utilisé ce terme pour exprimer leur doute quant à son statut de corpus. En tout cas il est clair qu'on y est confronté à un ensemble de textes dont la nature et la qualité ne se retrouvent dans aucun corpus constitué à partir d'autres sources. J'ai donc très rapidement dû mettre en place des procédures automatisées de filtrage pour traiter les différentes sources de bruit rencontrées dans les résultats bruts, c'est-à-dire en automatisant le téléchargement et l'analyse des documents pour en extraire l'unité lexicale et son contexte.

4.4.2.1 Sources d'erreurs communes

Les types d'erreurs suivants ont été identifiés et traités comme indiqué ci-dessous. Les choix parfois drastiques sont justifiés par une volonté de limiter le travail de dépouillement, et par le fait qu'étant donnée la masse, un mot légitime finira bien par y résister.

- **Absence de mot** correspondant au schéma. Entre le moment où une page est indexée par le moteur et celui de l'interrogation, plusieurs modifications ont pu entraîner l'impossibilité de trouver le moindre mot correspondant au schéma. La page a pu simplement disparaître ou son contenu a pu être totalement modifié par l'auteur. Certaines erreurs de segmentation en mots dans le processus d'analyse par le moteur ont pu également entraîner des erreurs (notamment lorsqu'une balise de formatage est utilisée en milieu de mot, par exemple pour des lettrines ou autres effets typographiques. Aucun traitement spécifique n'est à effectuer à ce stade.
- Le mot est en fait un **nom propre** (patronyme, toponyme, nom de marque, etc.). Pour éviter ces problèmes, seules les chaînes en minuscules sont retenues.
- **Erreur d'orthographe** ou de frappe. Cette source de bruit est une des plus importantes. L'éventail des types d'erreurs est très large, et une procédure de vérification automatique a été mise en place. Pour éviter de rejeter toute formation nouvelle (inconnue du lexique de référence utilisé par ce type d'outil), je me suis limité à la détection des erreurs suivantes :
  - les fautes d'accents, quelles qu'elles soient, et sans limite de nombre par mot, comme "*prêfèrable*" pour "*préférable*" ;

- les dédoublements (ou pire) de lettres comme "*grottesque*" pour "*grotesque*" ou au contraire : "*décolage*" pour "*décollage*" ;
- l'inversion de deux lettres consécutives comme "*obliagtion*" pour "*obligation*" ;
- l'ajout ou la suppression d'une lettre comme "*adapatable*" pour "*adaptable*" ou bien "*abillage*" pour "*habillage*" ;
- la modification d'une lettre comme "*mertion*" pour "*mention*".

– Vérification de la **segmentation des mots**. Il s'agit en fait d'une vérification orthographique en contexte destinée au traitement des mots collés ou mal découpés comme dans l'exemple ci-dessous où *avantageusesque* n'est pas un adjectif en *-esque* :

> *les prestations obtenues sont moins* ***avantageusesque*** *celles dont bénéficie un salarié à revenu égal,*

Dans ce cas, on rejette le mot s'il existe un découpage qui donne deux mots présents dans le lexique de référence ("*avantageuses + que*") et qui a une fréquence sur Altavista supérieure à celle du mot suspect : "*avantageuses que*" est présent dans 785 pages alors que "*avantageusesque*" ne l'est que dans une seule.
Cette correction peut cependant être à l'origine de découpages abusifs comme dans le cas de l'adjectif "*lestable*" qui peut être découpé en "*les + table*" ; or il n'y a que 105 occurrences pour l'adjectif alors qu'il y a 257 occurrences de la séquence erronée "*les table*". En d'autres termes, certaines fautes de frappe et d'accord surviennent plus fréquemment que certains mots construits...
Le problème inverse se pose dans le cas de textes gardant des traces d'une mise en page préalable à leur formatage HTML, comme les césures dans l'exemple ci-dessous. Dans ce cas, la présence d'un tiret à gauche du candidat "*mentation*" permet de vérifier la pertinence du recollage sur les mêmes principes que précédemment.

> *créer une réserve d'eau pour l'ali-* ***mentation*** *en eau potable de la région...*

– **Contexte dans une autre langue**. Altavista, comme tous les moteurs de recherche généralistes sur le Web, effectue un diagnostic de langue sur les pages qui ne l'indiquent pas explicitement dans leurs en-têtes. Dans un premier temps, les requêtes générées par Webaffix indiquent qu'on limite la recherche aux pages rédigées en français, mais cela n'est pas suffisant et des erreurs subsistent. Le problème se pose d'autre part pour les pages multilingues. Altavista n'attribue en effet qu'une seule langue à chaque page Web, a priori en fonction du début du document ou de la langue majoritaire. En résultat, la forme candidate peut très bien apparaître dans un segment en anglais ou en espagnol au sein d'une page par ailleurs en français.
Au bout du compte, Webaffix vérifie systématiquement, dans une fenêtre de 100 caractères autour de chaque occurrence du mot-cible, qu'il n'y a pas plus d'un mot-outil emprunté aux autres langues romanes et germaniques (anglais, allemand, espagnol, italien). Les mots-outils ont été sélectionnés par leurs fréquences, en enlevant les cas de recouvrement avec le français. Par exemple, "*or*" n'appartient pas à l'antidictionnaire de l'anglais.
Quelques problèmes résiduels demeurent, par exemple pour les segments trop courts comme ci-dessous, qui est un cas classique de citation d'un titre original :

> *il nous faut aller la trouver dans les pages de Sept jours pour expier (days of* ***atonement****) de WJ Williams.*

La méthode des mots-outils n'est pas non plus bien adaptée à la détection des langues trop proches, notamment l'ancien et le moyen français comme ici :

> *tant soit peu, diminue, Ny que ma foy descroisse* ***aulcunement****. Car ferme amour sans eulx est plus, que nue.*

– **Code informatique**. De nombreux contextes courants sur le Web sont les segments de code informatique, les URLs, les adresses mail, etc. qui peuvent contenir des chaînes de caractères correspondant au schéma recherché. La méthode de filtrage se fait simplement sur la base de certaines combinaisons de marques typographiques (notamment les slashes, les accolades, les soulignés...).

> *Method Summary (package private) void* ***actionaffichage****_détaillé() Méthode qui permet un affichage de...*
>
> *http ://www.****abacdepannage****.fr/*

A la suite de ces différents filtrages, les formes retenues présentent une précision autour de 40%. Le score est très variable entre les suffixes, tout comme varient énormément les sources de bruit (voir Tanguy et Hathout (2002); Hathout et Tanguy (2005) pour plus de détails). Les erreurs résiduelles concernent les cas les plus sévères des problèmes énumérés ci-dessus, des mots corrects mais ne correspondant pas au schéma dérivationnel visé, ou qui appartiennent à une autre catégorie grammaticale (adverbe au lieu de nom), mais aussi des situations plus complexes, qui correspondent cette fois au type de document rencontré plutôt qu'à la seule occurrence du mot visé. En voici les principales :

– **Textes générés automatiquement**. Nombre de textes rencontrés par cette méthode sont clairement issus d'un processus informatique et ne sont pas écrits par un scripteur humain. C'est bien entendu le cas de la traduction automatique, très couramment utilisée pour rendre multilingue à peu de frais un site Web. Dans certains cas cela entraîne le transfert direct d'un mot étranger dans un contexte considéré comme du français, mais également la formation de mots construits par des mécanismes apparemment intégrés à ces outils. Je prendrai comme exemple le mot *conservatricement*[14]. L'extrait suivant est sans contestation possible une traduction automatique (l'original contenait a priori l'adverbe anglais *conservatively*) :

> *Ils peuvent prévoir une réponse d'approximativement 35% des gens, à qui vous envoyez les E-MAILS. Mais nous laisser extrêmement* ***conservatricement*** *est tout d'abord et suppose que vous avez une quote-part moyenne de discours de seulement 10%. Si vous envoyez vos E-MAILS à 100 personnes différentes, vous pouvez prévoir que vous atteigniez au moins 10 de ces personnes, cela fait exactement cela, que vous avez fait.*

Comme on le voit clairement dans la dernière phrase, il s'agit d'une de ces *chain-letters* qui égayaient nos boîtes aux lettres avant les mornes spams actuels, et qui a atterri sur un forum. D'autres contextes de cet adverbe existent dans des textes plus honnêtes, toujours comme traduction.

– **Listes de mots et autres textes métalinguistiques.** Le Web regorge également de pages qui ne contiennent pas de texte rédigé, mais parfois des listes de mots n'ayant pas vocation à former des énoncés. Certaines pages sont ainsi conçues pour remplir la

14. Merci à Gilles Boyé pour cet exemple, ce type d'adverbes étant étudié dans (Plénat et Boyé, 2012, à paraître).

fonction d'*attrape-moteur*, de façon plus ou moins sophistiquée et attirer des internautes vers des sites plus ou moins recommandables. Des contextes comme celui ci-dessous étaient assez courants :

> *sites pirate sex passwords sexe argent blondes* ***naviguage*** *download proche orient palestine israel syrie naviguer*

Généralement les mots contenus dans ce genre de listes étaient empruntés à des textes « normaux », et dans ce cas le mot découvert (ici *naviguage*) était confirmé par d'autres occurrences (on trouve actuellement des centaines de contextes légitimes de ce dérivé, comme *Bon naviguage sur mon site !*). De plus, les systèmes d'indexation des moteurs de recherche semblent avoir fait beaucoup de progrès et évitent soigneusement ce type de document.
D'autres contextes amusants sont ceux où l'on tombe sur des travaux de linguistes, notamment lorsque ceux-ci déclarent un dérivé impossible :

> ***ridiculage*** *est morphologiquement impossible car on y reconnaît la base adjectivale ridicule (ridicule, ridiculiser, ...) et le suffixe -age, mais ce dernier doit se fixer à des bases verbales (laver -> lavage, couper -> coupage, ...) et non des bases adjectivales. (On pourrait par contre imaginer ridiculisage : le suffixe -is rend l'ajout de -age possible en transformant la base adjectivale en base verbale.)*

Bien entendu, le mot *ridiculage* est maintenant employé dans des dizaines de contextes tout à fait acceptables :

> *Tu vas illico poser une main courante au commisariat du coin pour tentative de* ***ridiculage*** *en public*
>
> *Pour t'éviter le* ***ridiculage*** *voire la ridiculitude, je suggère que tu parles de groupement*
>
> *Bah en même temps, j'ai un peu participé au* ***ridiculage*** *de Lmiara. On était d'ailleurs trois à se délecter de proses blairiennes*

– **Niveau de langue et compétence du scripteur.** Si des débats intéressants peuvent concerner l'acceptabilité des exemples précédents, nous avons toujours décidé de les considérer comme tout à fait légitimes, mêmes si certains effets contextuels sont à prendre en compte (voir plus loin). Par contre, dans certains cas il est clair que la circonspection est de mise, par exemple pour cette occurrence de *conservatricement* :

> *Pour résumer, je pense que Sohane n'est pas contente avec sa vie et cela est la raison principale, pourquoi je pense qu'elle doit se changer. D'après moi, des gens peuvent vivre* ***conservatricement****, s'ils sont heureux, mais Sohane a l'air d'être un peu jalouse de la vie de Djelila, sa liberté, ses amitié et son plaisir d'être en vie.*

Il s'agit ici d'un extrait d'un message de forum entre des lycéens allemands qui étudient le français (ici le message indique que c'est explicitement une demande de correction d'une rédaction : le titre est « *Korrektur s'il te plaît !* »). Pour l'histoire, un locuteur natif lui a proposé de remplacer *conservatricement* par *de façon conservatrice*

Tous ces exemples montrent bien la nécessité d'observer précisément tous les contextes d'apparition. Ce travail minutieux est toujours très riche d'enseignement, notamment dans les

exemples ci-dessus pour confirmer l'utilisation du procédé de dérivation (construire simplement l'adverbe sur la forme féminine de l'adjectif), que ce soit par des processus automatiques ou par des apprenants.

4.4.2.2 Analyse morphologique des dérivés

En plus du simple repérage de nouvelles formes suffixées, Webaffix comportait également un module permettant de calculer et de tester la forme de base supposée des candidats ainsi obtenus. Pour ce faire, deux phases supplémentaires étaient mises en place. La première correspond à un travail de calcul prédictif de la forme de base à partir du dérivé, en utilisant la méthode que Nabil Hathout avait mise au point pour la construction de ressources morphologiques (Hathout, 2000). La dernière étape consistait en une vérification du lien entre le candidat dérivé et la base prédite en recherchant des cas de cooccurrence des deux formes lexicales, là encore sur le Web. Ce module fonctionnait comme indiqué dans le schéma suivant :

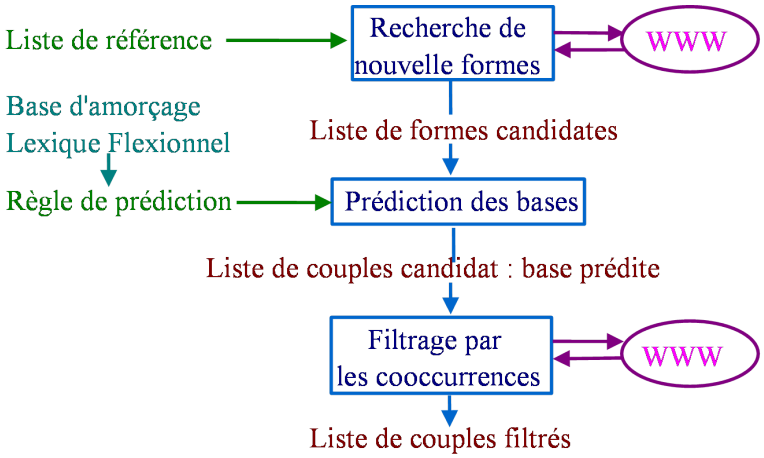

FIGURE 4.2 – Schéma de fonctionnement de Webaffix pour l'extraction de couples base-dérivé

Ainsi, l'analyse d'un nom déverbal comme *naviguage* arrivait à la prédiction du verbe *naviguer* comme base potentielle. Le couple *naviguage* - *naviguer* (en utilisant bien entendu des formes fléchies du verbe) était ensuite soumis comme requête, et les résultats vérifiés avec les mêmes procédures que pour le repérage initial. Le principe de validation d'un lien de dérivation par recherche de cooccurrence avait été découvert par Baayen et Neijt (1997), et mis en pratique immédiatement par Xu et Croft (1998) dans une application de recherche d'information.

Bien que constituant une contrainte exigeante, la procédure se révéla tout à fait efficace, et une quantité importante de bruit fut ainsi éliminée. Pour certains schémas dérivationnels (i.e. pour un suffixe comme *-age* et une catégorie de base donnée, par exemple le verbe) la précision finale pour les couples obtenus atteignait un très respectable 85%.

Ce type d'approche est, à mon avis, assez unique dans l'histoire du Web comme corpus, notamment par le soin apporté au filtrage des contextes. Les efforts généralement fournis dans cette mouvance concernaient, comme on l'a vu, la constitution d'un corpus le plus utilisable possible pour des études ultérieures, et ne pouvaient donc chercher à filtrer aussi précisément. De plus, le groupe Wacky s'est rapidement focalisé sur une tâche de nettoyage

qui ne présentait pas d'intérêt pour notre approche. La campagne *Cleaneval* (Baroni *et al.*, 2008) visait principalement à rechercher des duplications dans les pages accumulées, et surtout à supprimer d'un document complexe comme une page Web les segments comme les barres de navigation, les en-têtes et pieds de page, etc. On comprend bien l'intérêt de ces travaux, puisqu'ils sont nécessaires si l'on veut obtenir des fréquences fiables, ce qui reste une approche majoritaire dans la linguistique de corpus. Malheureusement, cela ne concerne que très peu nos besoins, focalisés sur des phénomènes rares et sans aucun appel à la notion de fréquence.

4.4.3 Principaux résultats obtenus

Pendant les quelques années de son existence, Webaffix a été fortement mis à contribution. Je récapitulerai ici les principales campagnes de récolte et les questions scientifiques abordées (un aperçu global est disponible dans Hathout *et al.* (2008)).

4.4.3.1 Suffixes *-esque* et *-este*

Il était logique d'utiliser Webaffix pour répondre en détail à la demande initiale qui a entraîné sa création. La base d'adjectifs en *-esque* de Marc Plénat a donc été complétée par le biais de cette méthode. Mais le fait le plus marquant de ce travail, et sans doute un des plus beaux travaux scientifiques auxquels j'ai pu participer est l'étude des dérivés en *-este*, présentée dans Plénat *et al.* (2002). En 1940, Edouard Pichon avait repéré dans un texte de Verlaine le dérivé *Silvio-pelliqueste* (construit sur Silvio Pellico), dont il avait supposé qu'il se substituait à l'attendu *silvio-pelliquesque* à cause de la dissonance entraînée par le redoublement du phonème /k/. Les recherches de Marc Plénat lui avaient permis de repérer 3 autres dérivés de ce type (et vérifiant l'hypothèse), malheureusement tous trois du même auteur, Frédéric Dard, dans ses romans de la série San-Antonio que M. Plénat avait repérée depuis longtemps comme une source abondante de phénomènes de ce type. Une fois Webaffix opérationnel, il a suffi d'une vingtaine d'heures pour repérer 7 autres dérivés du même type. De plus, en relançant la même recherche (cumulative) quelques mois plus tard, de nouveaux dérivés furent repérés grâce à l'évolution constante et exponentielle du Web. Le tableau 4.1 répertorie l'évolution de l'inventaire de ces formes. Comme on le voit, la conjecture énoncée par Pichon se voit totalement confirmée, sans contestation possible au vu de la quantité de dérivés répondant tout à fait à la règle qu'il avait induite d'un seul exemple. Nul doute qu'un nouvelle campagne ajouterait encore de nouveaux dérivés, mais l'objectif était atteint.

4.4.3.2 Suffixe *-able*

Une autre étude initiée par Marc Plénat et Nabil Hathout concernait cette fois le suffixe *-able* et est détaillée dans Hathout *et al.* (2004). L'objectif de cette étude était de confronter les théories existantes sur le fonctionnement de ce suffixe à une quantité de lexèmes plus importante que celle utilisée classiquement, à savoir les seuls dérivés répertoriés dans les dictionnaires. La phase de récolte automatisée par Webaffix a permis de former une liste de plus de 5 000 adjectifs alors que les dictionnaires les plus exhaustifs cumulés n'en contenaient que 1400. Ce travail de dépouillement très important (malgré les méthodes de filtrage mentionnées plus haut) a ainsi permis de revisiter la description de ce suffixe.

Notamment, la recherche d'attestations sur le Web a permis de mettre au jour des utilisations plus libres de ce suffixe, appliqué à des bases verbales comme c'est normalement le cas, mais avec une signification bien différente de la glose traditionnelle de ces dérivés (un nom

Source	Date	Dérivés	Nb
Pichon	1940	*silvio-pelliqueste*	1
Plénat	1997	*astraqueste, grandiloqueste, dingueste*	4
Webaffix (1recampagne)	2002	*dingueste, hageste, iagueste, langueste, marqueste, pragmatiqueste, punkeste, titaniqueste*	12
Webaffix (2^{e}campagne) complété par Plénat	2004	*algueste, bangueste, big-bangueste, blagueste, blogueste, borgueste, bouledogueste, bouygueste, cirqueste, darkeste, dukeste, fiasqueste, fliqueste, gagueste, gangueste, jack-langueste, haddockeste, loqueste, luna-parkeste, mandrakeste, orqueste, pâqueste, pétanqueste, planckeste, ringueste, rockeste, stringueste, swingueste, tankeste, tongueste, turqueste, zingueste, zoukeste*	44

TABLE 4.1 – Évolution de la liste de dérivés en *-este*

est *V-able* si on peut le *V*). Ainsi, pour reprendre l'exemple emblématique de *pêchable*, ont été trouvés comme noms recteurs (en plus des différents poissons) :

– des jours, saisons, des conditions météorologiques et autres périodes de temps (*pendant lesquelles* on peut pêcher) :

> *Ne parlons pas du mois d'août... Impêchable !*

– des rivières, étangs et autre plans d'eaux ou lieux (*dans lesquels* on peut pêcher) :

> *3 km de rives pêchables, bien aménagées pour le lancer*

– du fil, des cannes à pêches ou tout autre matériel (*avec lesquels* on peut pêcher) :

> *je remarque après quelques lancers (je pêche généralement à 40 mètres en étang) que mon nylon se met à vriller et devient impêchable.*

– des tailles de poisson (*que ceux-ci doivent atteindre* pour qu'on puisse les pêcher) :

> *La sur-pêche et le non respect de la taille pêchable en Guadeloupe a entraîné une forte régression de la population.*

De plus, de nombreux cas de constructions dénominales ont pu être repérés, élargissant grandement les séries isolées qui avaient pu être repérées. On voit donc que la plasticité du suffixe est bien plus grande que ce qu'un échantillon plus réduit de données permettait d'observer. Dans certains cas également cette enquête a permis de trouver des attestations déclarées préalablement impossibles.

4.4.3.3 Concurrence suffixale : le projet Wesconva

Dans le même esprit de vérification des hypothèses établies concernant le fonctionnement des suffixes, j'ai participé avec Georgette Dal, Nabil Hathout, Stéphanie Lignon, et Fiammetta Namer (qui l'a dirigé) à un projet (financé par l'ILF) nommé Wesconva (pour Web, Suffixation, Concurrence de déVerbaux d'Action) présenté dans Dal *et al.* (2004). Ce projet se proposait d'étudier la concurrence suffixale des déverbaux, en étudiant notamment le cas où un même verbe donne lieu à deux noms d'actions, l'un en *-age* et l'autre en *-ment*, et d'étudier les cas

où l'un ou l'autre était présent dans un dictionnaire ou n'apparaissait au contraire que sur le Web.

La méthode utilisée à été légèrement différente, puisque nous sommes cette fois partis exclusivement des verbes répertoriés dans le TLFi, avons généré les dérivés possibles en utilisant la méthode WaliM de Namer (2003). L'approche de Fiammetta Namer est en fait hypothético-déductive et fonctionne en sens inverse de Webaffix : à partir d'une liste de bases et d'un processus de dérivation, des candidats dérivés sont générés et recherchés tels quels via des requêtes automatiques sur le Web. Nous avons donc décidé de mélanger les deux approches, et d'utiliser les fonctions de filtrage de Webaffix pour améliorer cette phase de vérification.

Au final, nous avons retenu 1 150 verbes présentant cette concurrence, et pour lesquels seul un des deux dérivés est répertorié dans un dictionnaire par exemple :
amocher : *amochage* (dict.) et *amochement* (Web)
estropier : *estropiement* (dict.) et *estropiage* (Web)

D'un point de vue quantitatif, le suffixe *-age* est majoritaire dans les dérivés absents des dictionnaires (65% des couples). D'un point de vue qualitatif, on peut en déduire soit qu'en les créant, le locuteur entend instituer une différence par rapport aux dérivés en *-ment* correspondants, s'il les connaît, soit, s'il ne les connaît pas, que le suffixe *-age* fonctionne tendanciellement comme suffixe par défaut.

Dans 30% des cas, le nouveau dérivé apparaît dans un domaine que ne couvre pas le dérivé du lexique conventionnel (par ex., *décagement* : /ornithologie/, /mine/ vs *décageage* : /agro-alimentaire/). Dans 70% des cas, le nouveau dérivé n'est donc pas motivé par une différence de domaine d'emploi (ex. *gravillonnage*, *gravillonnement* : /équipement/).

Du point de vue des théories préalablement posées pour expliquer la distinction entre les deux suffixes, nous avons voulu tester les propositions de Kelling (2003). Son hypothèse est que la concurrence *-age/-ment* à base verbale constante est passible d'une explication en termes de proto-rôles : selon elle, le suffixe *-age* se combinerait avec des bases verbales dont le premier argument est proto-agent, tandis que *-ment* serait sensible, lui, à sa proto-patience. Par exemple, *battage* (*battage de tapis*) suppose un sujet qui effectue volontairement l'action, contrairement à *battement* (*battement de cœur*).

Sur nos données, les contextes révélateurs de la (proto-)agentivité du dérivé ont montré que les distinctions instituées par C. Kelling pour expliquer la concurrence *-age/-ment* ne correspondent pas à un phénomène nettement repérable en contexte. En effet, si 65% des déverbaux en *-age* ont au moins un emploi de type agentif, cela vaut aussi pour 46% des déverbaux en *-ment*.

Bien que les conclusions soient moins marquantes que dans les précédentes études, ce travail a montré que de tels phénomènes ne pouvaient plus être étudiés sans prendre en compte une plus grande variété de données. C'est notamment ce qu'a fait plus récemment Fabienne Martin dans (Martin, 2008), lorsqu'elle a étudié de nombreux dérivés repérés sur le Web pour proposer une distinction plus précise. F. Martin a d'ailleurs depuis utilisé des données obtenues par une adaptation de Webaffix dans (Martin, 2012, à paraître).

4.4.3.4 Lexique Verbaction

Dans une optique plus TAL que linguistique, Webaffix a également été utilisé pour constituer des ressources lexicales à large couverture. En l'occurrence, Nabil Hathout et moi-même

avons lancé plusieurs campagnes de récolte de noms déverbaux d'action suffixés en (*-ade*, *-age*, *-ance*, *-erie*, *-ement* et *-tion*) pour étendre le lexique Verbaction [15]. Ce lexique a été initialement conçu par Nabil Hathout à l'ATILF à partir des données extraites du TLFi et en utilisant la méthode de (Hathout *et al.*, 2002) et validées manuellement. Il contenait à ce stade 6 471 couples noms/verbes tels que le nom dénote l'action ou l'événement exprimé par le verbe. Par exemple, *élection/élire*. Après deux campagnes de collecte par Webaffix, en utilisant le module de prédiction et de vérification des bases (l'idée de cette technique est d'ailleurs issue de cet objectif précis), Verbaction contient actuellement 9 393 couples (donc une augmentation de 50% par rapport aux données initiales).

La couverture ainsi élargie permet de prendre en compte des couples correspondant à de nouveaux référents (*pacsage/pacser*), des termes techniques (*aquamarquage/aquamarquer*) ou des niveaux de langue (*baisage/baiser*) non couverts par la source d'origine.

Ce lexique est très utilisé pour un ensemble d'applications de TAL, comme l'analyse syntaxique ou l'extraction d'information.

4.4.4 Avatars de Webaffix dans l'adversité

Comme on l'a vu, les travaux d'exploitation du Web comme corpus sont soumis aux aléas des moteurs de recherche qui restent incontournables pour accéder aux données. Webaffix a donc subi comme d'autres un terrible coup lors du rachat du moteur Altavista par Yahoo ce qui a entraîné (au 1erAvril 2003) l'arrêt des requêtes par troncation : l'approche inductive utilisée jusque là ne pouvait donc plus être utilisée. Passée la période de deuil, nous avons envisagé deux types de réactions.

La première était de revoir à la baisse la couverture de la méthode, et de remplacer le premier niveau de Webaffix (voir figure 4.2) par une approche hypothético-déductive comme celle précédemment utilisée par Fiammetta Namer (Namer, 2003), c'est-à-dire partir de bases connues, construire des dérivés potentiels (en utilisant la technique de N. Hathout dans l'autre sens) et en interrogeant les moteurs de recherche avec ces formes complètes. Bien que rapidement opérationnelle (les deux autres niveaux restaient inchangés), cette méthode ne permettait absolument plus de répondre à des besoins comme l'extension de Verbaction, ni d'observer des créations simultanées de bases et de dérivés (comme par exemple le couple *wapiser/wapisable* construits tous deux sur l'acronyme *Wap*).

L'autre solution envisagée était, comme le prônait avec l'enthousiasme qui les caractérisent Kilgarriff et Grefenstette (2003) :

> « *This suggests a better solution : Do it ourselves. Then the kinds of processing and querying would be designed explicitly to meet linguists' desiderata, without any conflict of interest or "poor relation" role. Large numbers of possibilities open up. All those processes of linguistic enrichment that have been applied with impressive effect to smaller corpora could be applied to the Web. We could parse the Web. Web searches could be specified in terms of lemmas, constituents (e.g., noun phrase), and grammatical relations rather than strings. The way would be open for further anatomizing of Web text types and domains. Thesauruses and lexicons could be developed directly from the Web. And all for a multiplicity of languages.* »

C'est donc surtout Franck Sajous, alors récemment arrivé à l'ERSS qui a décidé de créer

15. Téléchargeable sur `http://www.univ-tlse2.fr/erss/ressources/verbaction/`

un moteur dédié au moissonnage de créations lexicales, nommé Trifouillette [16]. En lieu et place d'un moteur de recherche, Trifouillette était un *crawler* qui parcourait le Web (en suivant les liens hypertextes), mais en n'indexant que les pages contenant des mots nouveaux, et en mettant en place un système d'interrogation, d'alerte et de validation dédié aux travaux menés dans l'équipe. Malgré les louables efforts et la compétence de F. Sajous, ce projet a malheureusement dû être abandonné, face à la complexité de l'opération de parcours du Web (et de la ruse dont il faut faire preuve pour échapper aux multiples pièges tendus par des sites Web peu conformes aux normes édictées), des ressources matérielles exigées (notamment en termes d'accès réseau) pour arriver à atteindre un rythme de moissonnage suffisant.

La dernière solution était donc d'utiliser des corpus tout faits, comme ceux fournis par les travaux mentionnés en 4.2.3, ou encore mis à disposition (en fonction de leur bonne volonté) par des moteurs de recherche. Dans Hathout *et al.* (2009) nous montrons ainsi comment nous avons pu relancer une dernière campagne d'extension de Verbaction en utilisant un corpus fourni par la société Exalead (que je remercie une fois encore). Si là encore la plupart des procédures développées pour Webaffix peuvent être recyclées à peu près telles quelles, les résultats obtenus sur un corpus statique ont du mal à justifier l'effort fait pour s'y adapter : une fois exploité le corpus n'est plus utilisable. Par contre, cela nous a permis de mesurer précisément le volume nécessaire à l'acquisition des données du type de celles présentées dans la section précédente : le volume est absolument décisif, et rien ne peut remplacer le Web pour y chercher des phénomènes rares.

4.5 Quelques pistes à explorer

Malgré les déconvenues récentes et l'actuel manque de moyens permettant de relancer de grandes campagnes d'acquisition de données, les nombreuses expériences menées lorsque cela était possible ont ouvert un ensemble de questions de recherche que je vais exposer ici.

C'est sans doute moins le travail mené pour effectuer les opérations de filtrage que les dépouillements (ou dépouillages) manuels des données plus ou moins brutes qui ont permis d'identifier des phénomènes que j'estime intéressants à examiner.

Le premier point concerne la question du genre des pages Web, dont on a vu à travers notamment les travaux de Santini (2007) qu'elle était loin d'être résolue, puisque la diversité et la nature des genres du Web ne fait pas l'objet d'un consensus. Il est pourtant évident que des progrès dans cette direction pourraient bénéficier directement aux études des phénomènes cités dans ce chapitre, et à l'exploitation du Web comme corpus en général.

Il est par exemple clair que certains genres du Web sont plus productifs que d'autres, et se concentrer sur ces documents ferait gagner en efficacité (on a vu que M. Plénat avait repéré la grande productivité de San Antonio, et les morphologues de l'ERSS ont depuis quelques années identifié des forums Web comme *doctissimo* ou *aufeminin.com* comme étant des mines sans fond de créations suffixales innovantes.)

Certains genres sont également plus pertinents que d'autres en termes de stabilité des lexèmes qui y sont découverts, et certains contextes repérés à la volée sont vus avec moins de circonspection que d'autres par les linguistes qui les utilisent comme simples exemples. En ceci, nos travaux sur le repérage des néologismes se distinguent de ceux de Valette (2010) qui ont une visée lexicographique, et de ce fait tendent à rejeter certains genres du Web qui ne garantissent pas la moindre autorité éditoriale.

16. `http://w3.erss.univ-tlse2.fr/membre/fsajous/trifouillette/`

Pour l'instant les traits utilisés pour les approches en classification automatique (par apprentissage automatique supervisé, voir partie IV) sont :

- des traits structurels concernant l'organisation logique du document : sans doute les traits les plus productifs pour séparer les grandes catégories, comme on l'a vu dans les toutes premières explorations de la question, par exemple le projet TypWeb (Beaudouin *et al.*, 2001) ;
- des traits lexico-syntaxiques comme ceux utilisés dans le travail fondateur de Biber (Biber, 1988) ;
- des configurations ponctuationnelles qui ont été identifiés comme pertinents pour des distinctions spécifiques, comme par exemple la distinction entre sites Web racistes et antiracistes (Valette et Grabar, 2004).

Un phénomène qui mérite à nos yeux d'être examiné à large échelle est celui des *rafales suffixales*, autrement dit des séquences contenant des séries de termes suffixés (généralement hors dictionnaire). Ces rafales suffixales avaient été repérées initialement lors du projet Wesconva, et nous nous étions posé la question de la recevabilité de dérivés dont la création semblait répondre à un besoin très local et stylistique, en grande partie humoristique. Nous les avons au final écartés des données utilisées pour quantifier les phénomènes étudiés, mais précieusement gardés. En voici quelques exemples :

> *J'ai testé pour vous... le* ***visionnage*** *juste après* ***lecturage*** *!*
> `http://www.melonthecake.com/page/10/`
>
> *Hobbies :* ***Cuisinage****,* ***véloballadage*** *,* ***lecturage*** *,* ***cinéphage*** *...*
> `http://www.viadeo.com/fr/profile/`
>
> *Jeudi, nous débutons une journée classique dans une famille :* ***levage****,* ***douchage****,* ***mangeage*** *(pancake au miel).*
> `http://amelieetyoann.over-blog.com/article-j-93-a-j-99-79192223.html`

Ce type de rafales en *-age* est très fréquent dans les textes dont les auteurs racontent une série d'actions en insistant sur l'accumulation et/ou le côté rituel.

> *Puis* ***rangeage****,* ***nettoyage****,* ***vidage****,* ***goutage****,* ***siestage****,* ***douchage****,* ***mangeage****,* ***dormage****....*
> `http://www.sentiersnomades.com/spip.php?article73`
>
> ***Ramenage*** *d'enfant chez son père.* ***Douchage****.* ***Mangeage****. Et* ***attendage*** *de turquoise.*
> `http://inkr3dible.canalblog.com/archives/monsieur_turquoise_/index.html`

L'utilisation de déverbaux est normale pour une telle énumération d'actions, mais le recours à des dérivés hors dictionnaire est très frappante. Si le suffixe *-age* semble majoritaire, il n'est pas exclusif :

> ***douchement****,* ***mangement****,* ***filmement****...*
> `http://maison.et.travaux.du.nefast.over-blog.com/[...]carrelage.html`
>
> *je vaque à les petites occupations du matin (discussion avec Filip,* ***douchation****,* ***maquillation****,* ***habillation****,* ***coiffation****... bref que des choses follement intéressantes ...)*
> `http://irlandetrip.canalblog.com/archives/2007/07/16/5780915.html`

De même que certains cas les suffixes sont mélangés :

10h :retour chez Laura dans un état à peine croyable, ***douchation*** *par groupe de 2,* ***séchage*** *puis* ***grattage*** *d'habits salubres et de maillots de bain(ce fut un combat difficile).*
*10h45 :***Raccompagnage** *de Julie chez Margot, puis* ***faisage*** *de pâtes pour le miam-miam.*
`http://la-jettatura.blogspot.com/`

La volonté ludique est évidente, et parfois explicite :

En fin d'après-midi retour au camping. Puis : ***arrivage****,* ***douchage****,* ***préparage****,* ***mangeage*** *et départ pour THE CONCERT (j'ai pas trouvé de rime...)*
`http://salsahora.free.fr/La-Seyne-sur-Mer-Festival-Cubain.html`

En fait, de telles créations lexicales sont très souvent commentées par les scripteurs :

Toutes ces tragédies, en plus de vous donner une expérience assez solide en matière de rupture en tout genre, de ***pleurnichage*** *dans les bras de vos proches et de* ***ridiculage*** *(comment ça, ca n'existe pas ce mot)(m'en fiche !)*
`http://uneautrequemoi.20six.fr/uneautrequemoi/art/1091791/Ainsi-va-la-vie`

Une hypothèse à creuser serait le rôle de l'amorçage dans ce type d'énumération, lorsqu'une séquence naturelle de déverbaux (présents dans les dictionnaires) semble entraîner une contagion et la création de nouveaux dérivés pour compléter la série :

mes besoins :
nettoyage *de mes planches après scann*
rattrapage *de dessin*
lettrage *(mais là ya pas besoin de palette)*
colorimétrie de planche mais simple hein faut pas pousser !
essayage *de dessin direct à la palette*

Après ***farfouillage*** *et* ***lecturage*** *j'opterais (?) pour : Bamboo Fun Medium Pen & Touch*
`http://www.bdamateur.com/forum2/viewtopic.php?id=10500`

Niveau ***élégance prestance classance*** *et* ***distinctance****, je reste sur mes positions*
`http://ashleyandr.blogspot.com/2009/05/moda-vant-tout-le-monde.html`
(exemple emprunté à Dal et Namer (2010))

Certains cas sont également clairement parodiques, comme ceux qui se basent sur des créations très médiatisées. Par exemple le célèbre *bravitude* de Ségolène Royal a généré une quantité importante de dérivés parodiques en *-itude*[17] :

Ignioritude
7 décervelage, Sainte Forçats, pollorcètes
Hier, notre ***candidateuse socialistique*** *à la* ***présidenciation*** *de la* ***Républiquitude*** *était en* ***baladage*** *chez les Chinetoques.*
`http://www.melfrid.net/index.php?post/2007/01/07/130-ignioritude`

17. Le site `http://www.echolalie.org/wiki/index.php?ListedItude` en répertorie un très grand nombre.

Si la plupart des travaux sur ces dérivés éphémères se concentrent sur leur intérêt pour l'étude des mécanismes de la suffixation, ils n'utilisent le contexte que pour expliciter leur interprétation. Une étude à large spectre de ces contextes permettrait sans doute d'observer le phénomène plus globalement, et de délimiter les conditions de leurs formations.

On rejoindrait donc ainsi les travaux de caractérisation des écrits du Web évoqués en première partie de ce chapitre.

On voit donc à travers ces différentes approches la grande diversité des études possibles sur le matériau langagier issu du Web. La place des outils y est par contre différente de celle qu'ils prennent pour l'interrogation de corpus classiques. Certes, l'outillage informatique sert encore une fois à gérer la masse, mais cette fois il a un rôle plus proche de la nature des données, en s'attaquant au filtrage et à la caractérisation des contextes. Ce sont donc les données elles-mêmes qui posent au final le plus de problème, et cet état de fait a des implications bien au-delà des pratiques des linguistes (voir notamment les écrits collectifs de Pédauque (2003) sur les évolutions qu'a entraînées le développement du document numérique). Pour une discipline si attachée à son matériau, le manque d'informations, de stabilité, de fiabilité voire de matérialité des documents trouvés sur le Web sont autant de problèmes majeurs que la linguistique doit affronter de face.

Pour les écrits du Web comme pour les corpus plus traditionnels, les besoins sont spécifiques à l'étude envisagée, et le développement ne peut se faire indépendamment d'une compréhension fine des objectifs scientifiques, tout comme il ne peut reposer sur une solution logicielle générique préconçue. Même les gros corpus issus du Web ne répondent pas à tous les besoins que le Web a su faire émerger : aussi volumineux soient-ils, ils restent statiques.

Une des pistes les plus intéressantes concernerait à mon avis la mise en place de procédures automatiques de caractérisation à la volée, ne visant pas à la catégorisation en genres, mais permettant par contre de donner des informations utiles sur une page Web pour une exploitation linguistique (par exemple l'identification de contextes considérés comme licites). De telles procédures couvriraient des besoins en googleologie, et seraient insérables dans des approches plus lourdement outillées.

Troisième partie

Réduire la complexité des données : techniques de visualisation et analyses statistiques

Cette partie aborde le deuxième pan de mes activités autour des données langagières, qui concerne leur analyse et non plus seulement leur recueil. Que cela soit dû à la nature intrinsèque du langage ou au développement de modèles théoriques et d'outils d'annotation sophistiqués, les informations disponibles sur les données langagières atteignent de hauts niveaux de complexité, en plus, comme on l'a dit, de volumes croissants. L'informatique peut alors apporter des solutions sous la forme de modèles de structuration des données, mais aussi de représentation synthétique de celles-ci. Dans ce cas, ce n'est pas nécessairement la technique informatique elle-même (en tant que calcul) qui est mobilisée, mais les modèles et outils qu'elle emprunte aux mathématiques et aux statistiques et surtout met en œuvre.

Par rapport aux travaux présentés dans les deux chapitres précédents, ceux qui vont suivre correspondent à une implication différente de mon travail dans les travaux de recherche, puisqu'ils interviennent au cœur des questionnements scientifiques, et non plus simplement en amont. On verra que j'y ai plus clairement agi en tant que force de proposition, et pas seulement de réalisation, en définissant des méthodes d'analyse et de représentation constitutives du travail effectué, comme c'est souvent le cas, dans le cadre d'un projet collectif.

Cette partie est donc très importante dans la définition du rôle que je me suis attribué dans le domaine de la recherche en sciences du langage. De par ma culture scientifique initiale, je me sens plus à l'aise avec certains outils conceptuels pouvant être utilisés à profit dans des situations auxquelles est confrontée la linguistique. J'ai donc ainsi à plusieurs reprise eu la fonction de « passeur » interdisciplinaire en proposant des techniques classiques en informatique pour aborder des questionnements issus des sciences du langage. Ce passage est bien entendu toujours le lieu de compromis. On verra donc que les techniques de représentation et d'analyse que j'ai utilisées ne sont en aucun cas à la pointe des avancées scientifiques, et que les plus simples sont celles pour lesquelles l'interprétation est la plus facilement accessible. De même, dans certains cas la complexité des données langagières a volontairement été simplifiée pour s'adapter aux exigences techniques. On connaît bien la difficulté de ce type de dialogue interdisciplinaire, et je reste persuadé que les conditions d'une meilleure intégration passent à ce stade par une mise en avant des avancées réalisées sur des cas particuliers.

Dans le même ordre d'idée, j'ai volontairement multiplié les approches en choisissant des techniques différentes. Cette éclecticité est justifiée à la fois par la spécificité des données abordées lors de chaque travail particulier, et par la volonté farouche de ne pas disposer d'un seul outil supposé universel, situation dont on connaît la conséquence pour celui qui le tient à ne voir qu'un seul aspect des problèmes à traiter.

J'ai choisi de découper cette partie en deux chapitres, bien que comme on le verra ce sont parfois des questions similaires qui sont posées aux mêmes données. Il s'agit dans le cas général de l'observation, plus ou moins guidée par une hypothèse préexistante, de données langagières résultant d'un calcul ou d'une annotation. Le premier chapitre aborde une facette importante de mon travail que j'ai regroupée par le fait qu'elle m'a amené à chaque fois à proposer une représentation graphique pour visualiser les données. Le second effectue un tour d'horizons des quelques techniques statistiques que j'ai pu acquérir et déployer.

Chapitre 5

Visualiser les données langagières

Tous les manuels de statistique, qui proposent des mesures et des méthodes parfois complexes pour analyser des données insistent sur la nécessité de commencer par une phase de représentation graphique avant de déployer une technique d'analyse mathématique. Ce passage par une visualisation est en effet souvent bien plus éclairant sur la nature des différentes caractéristiques étudiées et les interactions qu'elles entretiennent. Mazza (2009) positionne cette activité dans le travail empirique comme un mode de passage des données vers l'information et la connaissance. C'est donc cet aspect que je vais montrer au travers de différents exemples de mon travail, que j'ai regroupés en deux catégories.

Le premier type concerne la représentation synthétique d'annotations, portant sur un ensemble complexe de données. Si, bien souvent, il ne s'agit pas d'une modélisation en tant que telle (au sens où le but visé n'est pas la prédiction), cela m'a souvent amené à utiliser des modes de représentation comme les graphes et les treillis, outils classiques issus des mathématiques discrètes.

Le deuxième type concerne plus spécifiquement le matériau textuel, dans des travaux qui visent cette fois à caractériser des phénomènes dans leur dimension longitudinale au fil d'un texte long. Si, là encore, ce sont les aspects de représentation visuelle que j'ai privilégiés, ce type de travail fait appel à des modèles conceptuels plus simples que les précédents, mais dont l'exploitation est prometteuse. Je me contenterai ici de montrer en quoi la simple représentation graphique permet d'identifier des structures particulières dans un texte et de donner une première réponse aux questionnements sur ce type de données.

Dans une dernière partie, je vais tâcher de dégager un ensemble de pistes de réflexion sur la place de ces représentations visuelles dans l'exploration et l'étude de données langagières. Ce recul prendra appui sur un ensemble de travaux plus généraux sur la visualisation de données, formant un champ disciplinaire récent, pluridisciplinaire et très dynamique, qui profite pleinement des avancées technologiques des outils informatiques pour aborder le problème omniprésent de la confrontation, quel que soit le domaine étudié, à la multiplication exponentielle des données disponibles.

5.1 Synthétiser les données collectées

Certaines informations issues des données langagières sont trop complexes à analyser pour une approche directe. Que celles-ci soient recueillies par des moyens automatisés ou issues d'une opération manuelle d'annotation ou d'extraction, elles forment une collection qui se

présente comme un amas de données individuelles. L'obstacle principal à leur exploitation n'est pas nécessairement celui de leur masse, mais de la complexité de leur organisation et de la variété des modes de classement.

Comme on le verra dans les travaux spécifiques que je détaillerai, il s'agit la plupart du temps d'une approche exploratoire. Cela ne signifie pas que les objets langagiers étudiés ne bénéficient pas d'un apport théorique ou empirique préalable concernant leur caractérisation, mais plutôt que le travail n'implique pas la formulation d'une question précise, comme la vérification d'une hypothèse. Ce genre de questions ouvertes est courant en linguistique empirique, quand on cherche par exemple à caractériser un corpus particulier, ou à étudier les différents contextes d'une unité lexicale ou d'une structure phrastique. Si dans ce domaine particulier on a vu se forger au fil des années des méthodologies et des outils (notamment statistiques) permettant d'aborder l'exploration en bénéficiant d'un certain nombre de techniques éprouvées, les données que j'ai rencontrées ne rentraient pas toujours dans des cadres balisés, et l'innovation était alors nécessaire, pour le pire et le meilleur.

Sur la base de mon expérience face à des situations très variées, j'ai pu identifier les trois grands types d'objectifs suivants que doit tenter d'atteindre un travail d'analyse de ces données :

- *avoir un point de vue global* sur une relation établie entre des objets, pour en présenter la complexité et observer des configurations locales particulières ;
- *rechercher des configurations* récurrentes dans l'ensemble de ces données, et en déduire des comportements stables ;
- *croiser des données de nature différente* pour envisager leur interaction.

5.1.1 Avoir un point de vue global sur des relations isolées

Les deux exemples que je vais détailler ici se caractérisent par une méthode commune, mais se situent à la marge de la linguistique. Les données exploitées proviennent dans les deux cas d'une analyse basée sur un matériau langagier, mais sans aborder des questions portant directement sur le langage. Cependant, l'éclairage que mes propositions ont pu apporter peut tout à fait être appliqué à des matériaux plus classiques en analyse de corpus.

5.1.1.1 Le schéma du délire interprétatif de Madame M.

Comme indiqué dans la première partie (section 1.1.4, page 31), le travail que je présente ici s'est fait en collaboration avec Michel Schmouchkovitch, médecin psychiatre à l'hôpital de Brest, et mon encadrant de thèse Ioannis Kanellos ; le contexte et l'étude sont présentées dans (Schmouchkovitch *et al.*, 1998).

Cette étude très particulière concerne Madame M., une des patientes que suivait à l'époque M. Schmouchkovitch, et qui présentait une pathologie mentale lourde (psychose schizophrénique). Une des manifestations notables est apparue au cours des entretiens avec son psychiatre sous la forme d'un code spécifique qu'elle avait établi, et qui lui permettait d'interpréter les éléments langagiers et numériques de son environnement, en associant à chaque lettre et chiffre une signification particulière, apparemment stable. C'est pour définir et étudier en détails ce système interprétatif que M. Schmouchkovitch nous avait proposé de travailler avec lui sur la base des entretiens (qu'il enregistrait avec l'accord de la patiente).

J'ai proposé que soit effectué tout d'abord un recueil systématique de tous les éléments du code de Madame M., en lui demandant d'expliciter le sens exact attribué à chaque symbole,

et la motivation de celui-ci. Cela s'est fait par étape, sur plusieurs séances, avec une discussion et une analyse entre chacune pour orienter l'investigation et dégager des hypothèses à tester. Si les intuitions initiales de M. Schmouchkovitch étaient qu'il s'agissait de quelques éléments isolés, nous avons pu vérifier que ces interprétations étaient effectivement stables, et qu'elles formaient bien un système, tel qu'il est rendu dans le schéma de la figure 5.1. Dans ce schéma, les signifiants sont indiqués en gras et les signifiés en caractères normaux ; les relations sont étiquetées pour indiquer le type de motivation et les rapprochements entre signifiants ou entre signifiés.

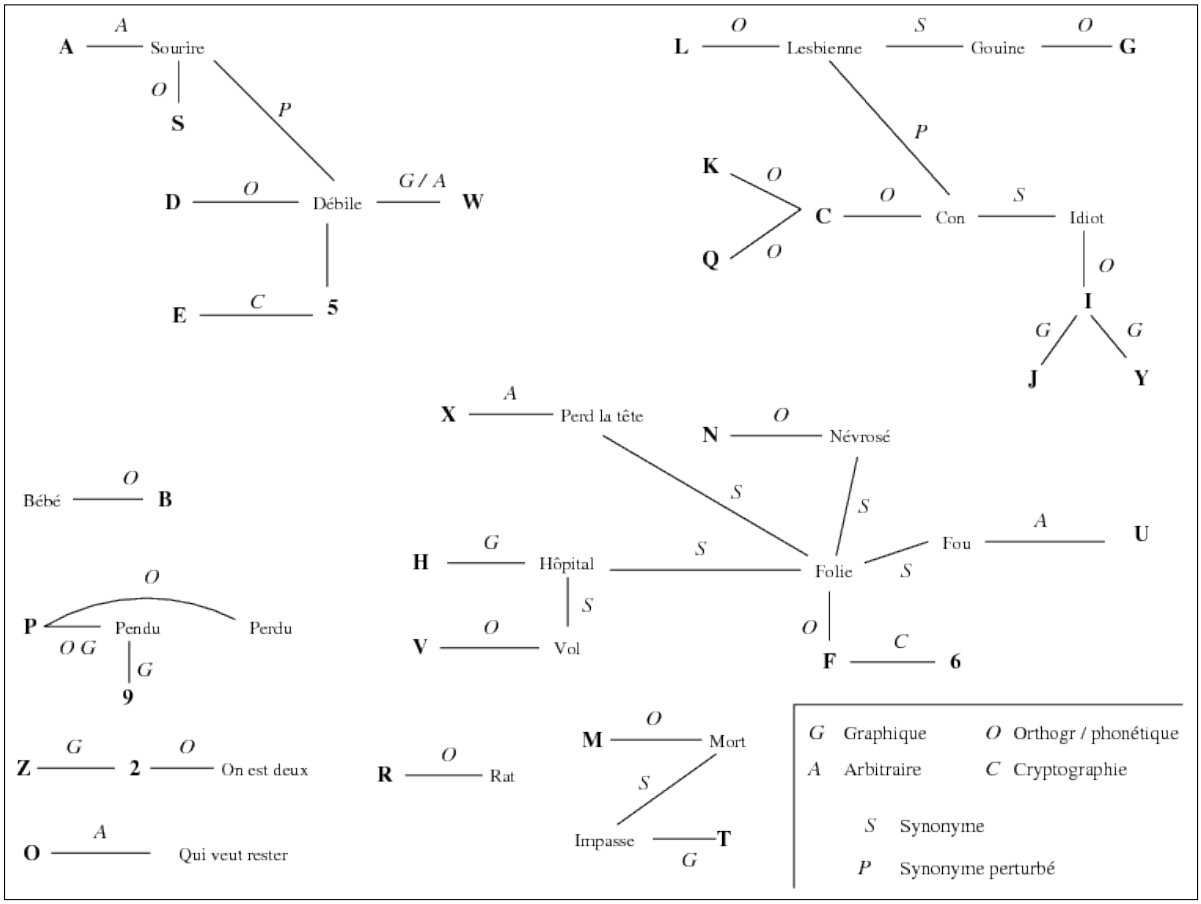

FIGURE 5.1 – Graphe du code interprétatif de Madame M.

Les concepts identifiés dans ce code sont bien entendu bien plus complexes que les seules expressions utilisées pour les noter (la notion de *rat* par exemple était glosée par la patiente comme *quelqu'un qui est mal dans sa peau* et n'évoquait pas spécialement un rongeur ni les autres acceptions figurées de ce mot). On voit ainsi se former au fil de l'analyse des rapprochements assez complexes entre des notions et des formes, qui sont rendus visibles sous la forme de sous-graphes connexes. Plusieurs expériences ont montré (une fois le graphe établi) que la patiente était tout à fait capable de confirmer l'existence de liens transitifs à longue distance entre des éléments de signification qu'elle avait explicités individuellement.

L'établissement de ce schéma avait donc deux intérêts :

– permettre d'identifier des composantes complexes où des concepts apparemment différents se révélaient être voisins ou assimilés, et observer les points focaux du délire de Madame M. (et surtout les distinctions entre ces composantes, qu'elle se refusait totalement à faire, par exemple entre *idiot* et *débile*, deux notions absolument distinctes pour elles) ;
– guider le « recueil des données » en organisant les questions au fil des entretiens, comme on l'a noté ci-dessus.

Ce travail a bien entendu été guidé par la cadre théorique de la sémantique interprétative dans lequel je me situais pendant mon doctorat. C'est notamment la recherche de traits sémantiques partagés par les différents signifiés qui a guidé à la fois l'investigation et la représentation des données.

Notons qu'aucun processus automatisé n'a été utilisé pour ce travail, mais c'est simplement ma culture d'informaticien de l'IA qui m'avait conduit à utiliser ce type de représentation pour traiter cette question. La représentation graphique d'un réseau de relations reste toujours un excellent outil de synthèse, facilement lisible et directement utilisable, comme on va le voir dans un second cas d'utilisation.

5.1.1.2 L'enchaînement des thèmes dans une consultation médicale

Le projet Intermède, évoqué en 2.3.3, a été une source importante de données riches et variées, que l'on verra à différentes reprises dans ce mémoire. Ce projet avait pour objectif d'articuler des recherches pluridisciplinaires autour de la question des inégalités sociales de santé, et mobilisait des épidémiologistes, des médecins, des sociologues, des psychologues et des linguistes autour d'un matériau commun : des retranscriptions de consultations médicales. L'objectif principal était de regarder, par différentes méthodes, si des inégalités sociales pouvaient être observées au sein de cet échantillon de plusieurs dizaines de consultations.

Une des approches a consisté en une étude systématique des sujets abordés dans le dialogue entre le patient et le médecin. Sur une quarantaine de consultations retranscrites, les sociologues membres de l'équipe SOI (Sports, Organisation, Identité, EA 3690, Université de Toulouse 3) Jean-Paul Génolini et Roxane Roca ont effectué manuellement une segmentation en passages et un étiquetage thématique. Si leur préoccupation principale concernait les sujets liés à la prévention des risques cardio-vasculaires (alimentation, suivi pondéral, consommation d'alcool et de tabac, activité physique), ils ont attribué une classe thématique à chaque passage dialogique afin de mettre au jour des profils génériques de consultation (Génolini *et al.*, 2011). Aidés dans cette approche par le logiciel Modalisa [1] qui propose une assistance pour cette tâche (généralement dans le but d'aider à l'analyse d'entretiens et d'enquêtes), ils souhaitaient obtenir une représentation plus globale des enchaînements des différentes thématiques. Le logiciel utilisé propose un ensemble de fonctionnalités pour exploiter ce type de données, dont des analyses statistiques sur la fréquence de chaque thématique (et la possibilité de croiser ces informations avec des variables caractérisant les différentes consultations), ainsi que la fréquence des transitions entre chaque thématique. Par contre, aucune possibilité n'était offerte pour une observation globale.

Je leur ai donc proposé une représentation graphique construite à partir des données qu'ils avaient annotées, en calculant la probabilité de transition entre chaque couple de catégories (donc en considérant l'étiquetage comme une chaîne de Markov). Le schéma résultat est présenté en figure 5.2.

1. `http://www.modalisa.com`

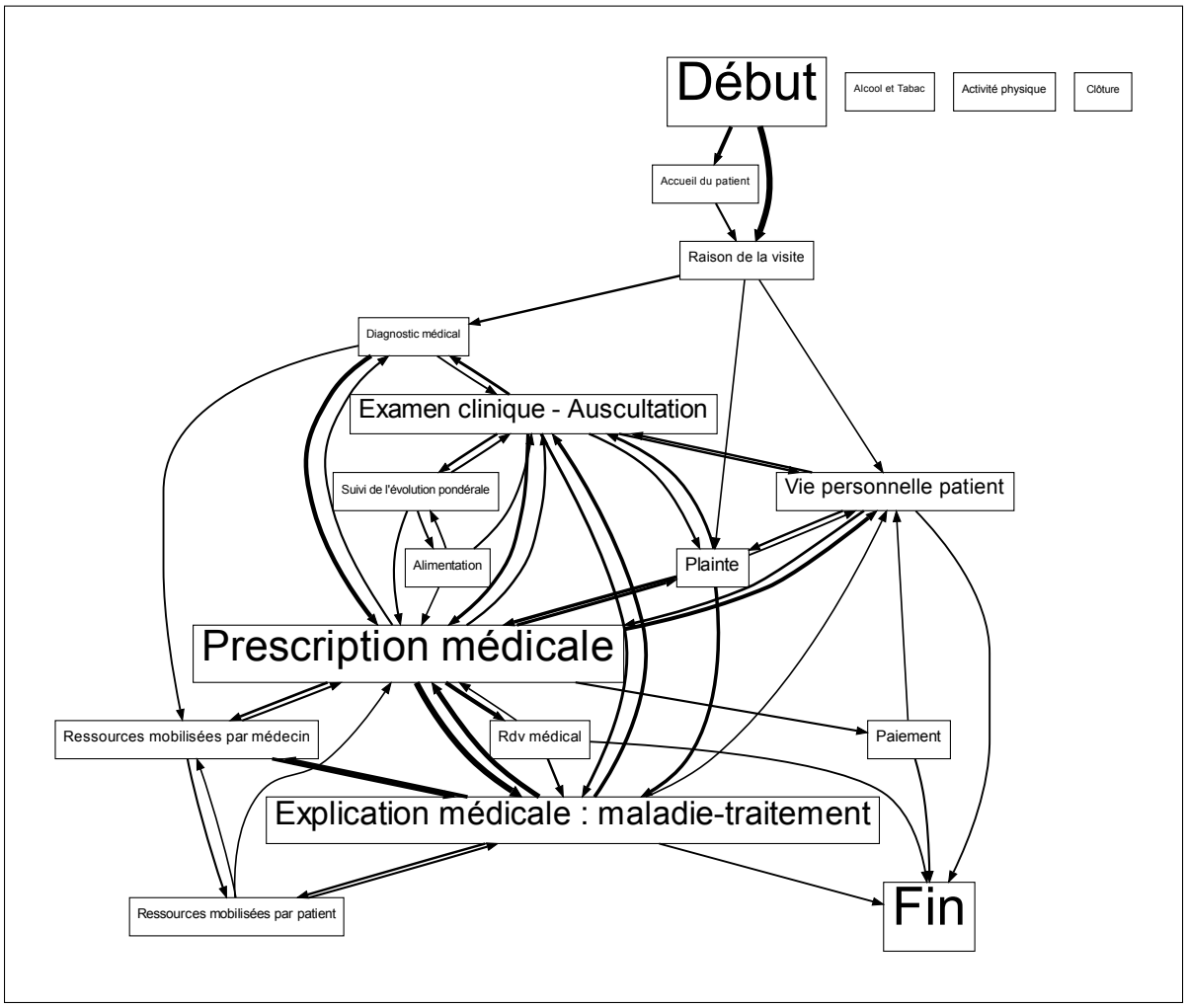

FIGURE 5.2 – Graphe de transition des thématiques dans les consultations médicales

Dans ce graphe les nœuds représentent les catégories thématiques, avec une taille relative à leur fréquence dans le corpus. Les arcs indiquent l'existence d'une transition observée entre deux catégories (dépassant un certain seuil de probabilité), en utilisant une épaisseur de trait proportionnelle à la probabilité de transition. La disposition planaire du graphe a été laissée aux bons soins du logiciel GraphViz [2], un allié précieux pour ce type de tâche. Si certaines catégories sont trop sporadiques pour être lisibles (certaines sont mêmes isolées par le seuillage, comme on le voit en haut à droite), les thématiques centrales sont, elles, bien identifiées. Voici le commentaire qu'en font J.-P. Génolini et R. Roca (extraits du rapport Intermede) :

> « *Le graphe montre que la consultation entre le début et la fin est organisée assez logiquement autour de la prescription médicale qui représente le nœud de l'activité médicale autour duquel se déploient les autres séquences de communication (principalement l'examen clinique, l'explication de la maladie, la vie personnelle du patient).*

2. `http://www.graphviz.org`

L'activité médicale sur l'éducation du patient se présente comme une pratique réflexive principalement articulée autour de l'alimentation et qui se loge généralement entre un diagnostic ou une prescription (raisonnement médical) en ouvrant des registres d'exploration particulièrement à partir du contrôle du poids. Lors des séances le médecin agit à l'intérieur d'un cadre délimité et opère des déplacements dans ses activités de prospection. Il peut, dans le moment de l'auscultation et de la surveillance pondérale interrompre momentanément ou définitivement cette conversation afin de porter attention à une douleur et se concentrer sur un symptôme. Dans l'économie générale de la consultation l'hygiène apparaît toutefois comme un module de communication cadré par un raisonnement médical classique sans lien de continuité avec les ressources des patients ou une approche de sa vie personnelle. »

On voit bien l'intérêt de cette représentation, tant du point de vue de l'organisation globale que de la focalisation sur des thématiques précises. Par exemple, la place centrale de la prescription apparaît clairement dans le graphe, en tant que thème structurant, entretenant des relations de contiguïté avec pratiquement toutes les autres phases du dialogue médecin-patient. La disposition globale, si elle a été décidée par l'algorithme de dessin, traduit bien clairement les grandes étapes (de l'exposé de la plainte jusqu'au paiement), qui ne transparaissent pas aussi clairement dans une table de transitions entre ces différents thèmes. Par contre, il est clair que l'examen local d'une articulation précise entre deux ou trois thèmes est plus facile en accédant directement aux fréquences de transition sous forme de tableaux. Cette complémentarité rejoint parfaitement les questionnements de Henry et Fekete (2008) sur les modes de représentations des réseaux sociaux, attribuant aux graphes une capacité globalisante mais rapidement ininterprétable lorsque la masse de données s'accroît, et qu'un zoom est nécessaire après identification d'une zone à examiner de plus près.

Si là encore le travail sur les données langagières a été fait en amont (et manuellement comme dans l'exemple précédent), on voit que mon rôle s'est résumé à la proposition d'un mode de synthèse nécessaire pour pouvoir exploiter les informations résultant de l'extraction des transitions. Un point intéressant ici est le fait que ce genre d'analyse ne faisait pas partie des fonctionnalités de l'outil utilisé (ModaLisa). Si dans certains cas les utilisateurs ont pu utiliser divers truchements pour combler certaines lacunes (par exemple utiliser les fonctionnalités d'un traitement de texte pour calculer le nombre de mots correspondant à chaque catégorie thématique), dans certains autres cas le fait d'entrer dans un cadre méthodologique objectivé par un outil multifonctions aussi sophistiqué soit-il est un obstacle à des approches innovantes.

5.1.2 Identifier des configurations : combinaisons d'indices des structures énumératives

Revenons maintenant plus au cœur de la discipline avec le projet Annodis et sa campagne d'annotation de structures discursives. Comme nous l'avons déjà présenté en première partie (cf 2.2.3), ces objets textuels ont été marqués manuellement dans le corpus en bénéficiant d'une assistance doublement outillée : par l'utilisation de l'interface dédiée Glozz (Wildlöcher et Mathet, 2009), et par le prémarquage automatisé de différents types d'indices qui, présentés à l'annotateur *via* l'interface, lui permet d'identifier plus rapidement et plus efficacement les zones du textes susceptibles de contenir de telles structures (Péry-Woodley *et al.*, 2009). Il faut noter à ce sujet que l'outil d'annotation Glozz constitue un effort remarquable dans une

direction similaire à celle que j'évoque dans ce chapitre, puisqu'il permet une visualisation interactive et à grain multiple de la disposition dans le texte des structures discursives.

Mais je vais me concentrer ici sur un autre aspect des données annotées : le projet Annodis se donne aussi pour objectif d'identifier la nature des différents éléments textuels aptes à signaler la présence d'une structure discursive. Dès lors, ces indices deviennent les objets d'une investigation, au même titre que les structures annotées elles-mêmes (dont les principales caractéristiques ont été décrites dans (Ho-Dac *et al.*, 2010)).

L'essentiel de nos efforts [3], a visé l'analyse des structures énumératives, dont le projet Annodis a permis de faire émerger une typologie, et d'en démontrer à la fois la fréquence et la variété dans les textes expositifs. La table 5.1 en montre quatre exemples tirés de la Wikipedia, qui correspondent aux quatre types proposés après observation du corpus annoté.

Dans l'exemple de la figure 5.3, les segments de textes colorés correspondent aux éléments prémarqués de différents types. Certains de ceux-ci vont être sélectionnés par les annotateurs en même temps que les structures énumératives (ci-après SE) elles-mêmes, et leur être associés en tant que descripteurs.

Au final, chacune des 800 structures énumératives annotées s'est donc vue attribuer des indices de différents types, parmi lesquels :

- pour l'amorce (le segment introductif optionnel d'une structure énumérative) : des indices lexicaux (adjectifs numéraux, démonstratifs, des adjectifs introductifs comme *suivants*, etc.) ou ponctuationnels (essentiellement des deux-points) ;
- pour les items eux-mêmes, des séquenceurs ou des marqueurs d'intégration linéaire (*premier*, *deuxième*, *suivant*, *autre*, etc.), des groupes adverbiaux circonstants (temporels ou spatiaux), des structures syntaxiques parallèles ou des indices ponctuationnels (puces ou tirets introductifs, points-virgules, etc.) ;
- pour la clôture (le segment conclusif optionnel), des indices lexicaux similaires à ceux de l'amorce (mais orientés vers l'amont, comme *précédents*).

S'il est simple de calculer la fréquence d'association de tel type d'indice sur l'ensemble de la collection de structures énumératives, ou même d'identifier des préférences entre un type d'indice et un type de structure, l'identification de certaines configurations particulières est plus complexe. Une telle configuration peut être vue dans une première approximation comme la cooccurrence d'indices de types différents, qui sont observés avec une fréquence minimale. De telles configurations, en plus d'éclairer la connaissance des caractéristiques de l'insertion d'une structure dans le texte, peuvent également être utilisées comme des aides plus efficaces au repérage de structures, voire à terme pour l'automatisation de leur identification dans le texte.

L'observation de ces cooccurrences a, là aussi, nécessité une représentation spécifique. Avant tout, il est important de bien comprendre que les données présentent une très grande variété de cas : si certaines structures n'ont qu'un seul indice associé, certaines en cumulent jusqu'à 5 types différents. De plus, si l'on considère une combinaison d'indices comme la présence d'un indice lexical dans l'amorce et la présence de circonstants dans les items, il faut à la fois considérer sa fréquence isolée (le nombre de structures qui n'ont *que* ces deux types d'indices) et les cas où ils sont, plus ou moins systématiquement, accompagnés d'autres types. Cette situation m'a conduit à choisir une représentation en treillis pour la visualisa-

3. Ce projet a impliqué de nombreux chercheurs de CLLE-ERSS : Cécile Fabre, Lydia-Mai Ho-Dac, Marie-Paule Péry-Woodley, Josette Rebeyrolle et Franck Sajous pour la partie que je présente ici.
Le corpus Annodis est disponible à l'adresse suivante : `http://redac.univ-tlse2.fr/corpus/annodis/`

SE de type 1 (sections titrées) :

[6. Les conquêtes amoureuses de César]$_{Amorce}$
[6.1 *Les femmes de la haute société romaine*
D'après l'historien latin Suétone, César séduit de nombreuses femmes...]$_{Item\ 1}$
[6.2 *Les reines*
César a des relations amoureuses avec Eunoé, femme de Bogud,...]$_{Item\ 2}$

SE de type 2 (liste à puces) :

[Parmi les principaux organismes de normalisation-standardisation mondiaux, citons :]$_{Amorce}$
[- l'ETSI : European Telecommunication Standards Institute ou Institut européen des normes de télécommunication ;]$_{Item\ 1}$
[- l'ITU : International Telecommunication Union ou Union internationale des télécommunications ;]$_{Item\ 2}$
[- l'IETF : Internet Engineering Task Force ;]$_{Item\ 3}$
[- l'ATM Forum ;]$_{Item\ 4}$
[- l'ANSI : American National Standard Institute ;]$_{Item\ 5}$
[- l'IEEE : Institute of Electrical and Electronics Engineers.]$_{Item\ 6}$

SE de type 3 (multiparagraphique) :

[Une première observation est à faire à ce niveau...]$_{Item\ 1}$
[Une deuxième observation concerne le niveau où s'exerce la critique...]$_{Item\ 2}$
[Nos deux observations tirent dans la même direction]$_{Clôture}$

SE de type 4 (intraparagraphique) :

[Or la pertinence de telles délimitations est toujours à relativiser :]$_{Amorce}$ [à quel moment, par exemple, séparer l'Antiquité tardive du Moyen Âge ?]$_{Item\ 1}$ [Faut-il présenter l'art de l'Égypte ptolémaïque aux côtés de celui de l'antiquité grecque ?]$_{Item\ 2}$
[Ou encore, si l'on convient de considérer la poésie comme un art, faut-il ou non présenter les poèmes de Léopold Sédar Senghor du côté des arts africains ?]$_{Item\ 3}$

TABLE 5.1 – Exemples de structures énumératives de chaque type

tion. Un tel treillis est présenté dans la figure 5.4 pour un type de structures énumératives (type 3 : structures énumératives autres que les listes et les sections, mais couvrant plusieurs paragraphes).

Un treillis est une structure mathématique correspondant à la représentation d'un ensemble ordonné, et qui est très couramment utilisée dans la représentation des connaissances en IA (d'où ma familiarité avec elle, acquise notamment pendant mon DEA). Chaque nœud du graphique représente une combinaison d'indices (la borne supérieure indiquée par *nil* représente l'absence d'indice). Plus un nœud est situé bas, plus il correspond à une combinaison spécifique, c'est-à-dire qu'il correspond à une configuration qui comporte un type d'indice de plus que les nœuds du niveau supérieur auxquels il est relié.

La taille des nœuds est proportionnelle au nombre de structures ayant cette combinaison

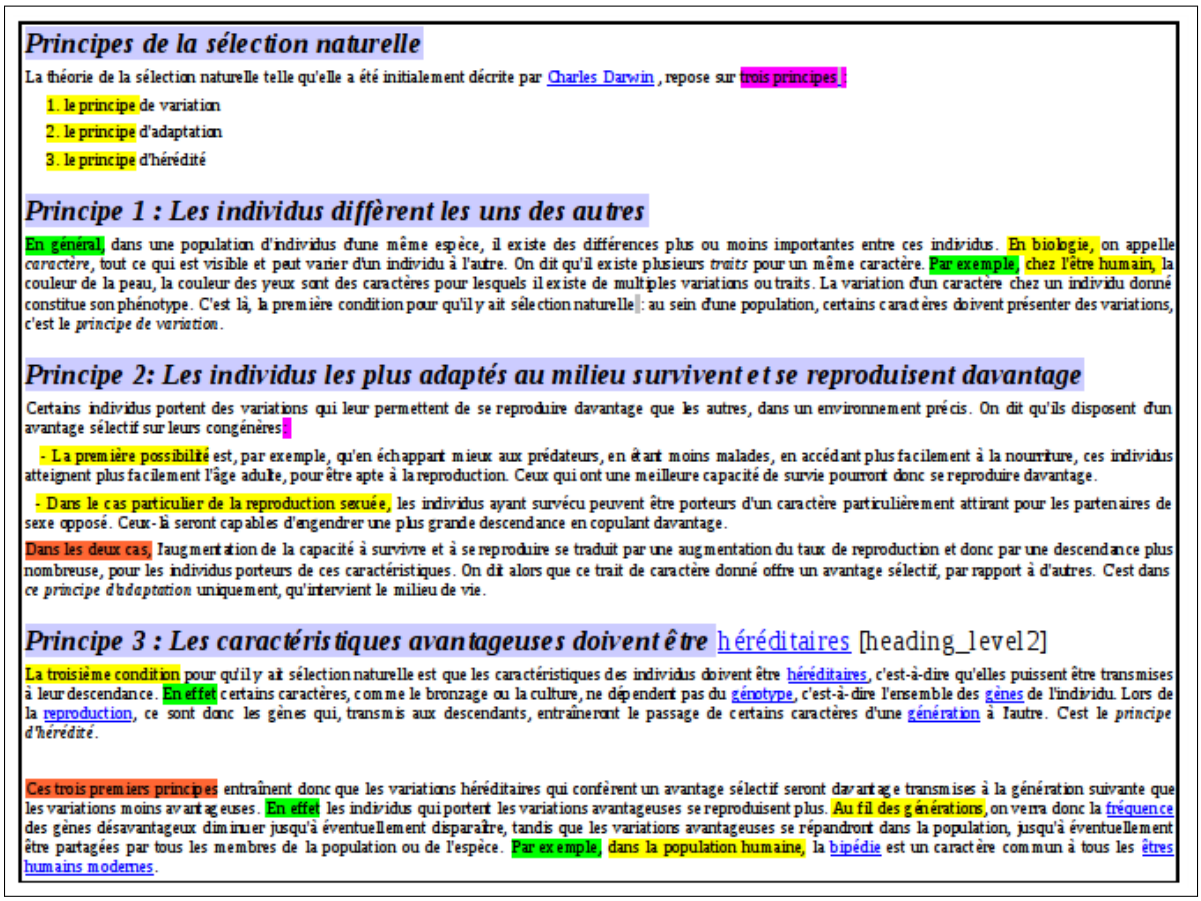

FIGURE 5.3 – Exemples d'indices prémarqués de structures énumératives
En fuchsia, les indices d'amorce – en jaune, les indice d'items dédiés (comme les marqueurs d'intégration linéaire) et les circonstants – en orange, les indices de clôture – en vert, les connecteurs – en gris, les titres de section

d'indices, et l'épaisseur des arcs correspond à la proportion de structures qui ont une combinaison d'indices plus spécifique. Par exemple, un arc très épais relie *Trigger Punct* (indice ponctuationnel d'amorce, premier nœud de la deuxième rangée) à *Trigger Punct + Trigger Lex* (indices ponctuationnels et lexicaux dans l'amorce) : cela signifie qu'une proportion très importante des structures qui ont un indice ponctuationnel dans leur amorce ont également un indice lexical. Pour aider à la lecture, chaque nœud indique deux pourcentages : le premier est simplement la part de structures qui ont cette configuration (au moins) et le second la part de celles-ci qui ont exactement cette configuration, sans indice supplémentaire.

Malgré la difficulté à lire ces données (mais on finit par s'y faire), plusieurs conclusions peuvent en être tirées, parmi lesquelles :

- On voit que pour les structures de ce type la présence d'un indice lexical dans l'amorce est une configuration très fréquente (nœud le plus grand).
- Les circonstants (*Circ.*) et les marqueurs d'intégration linéaire (ou séquenceurs, *Seq.*) sont utilisés dans une proportion de même ordre, mais par contre, là où les circonstants

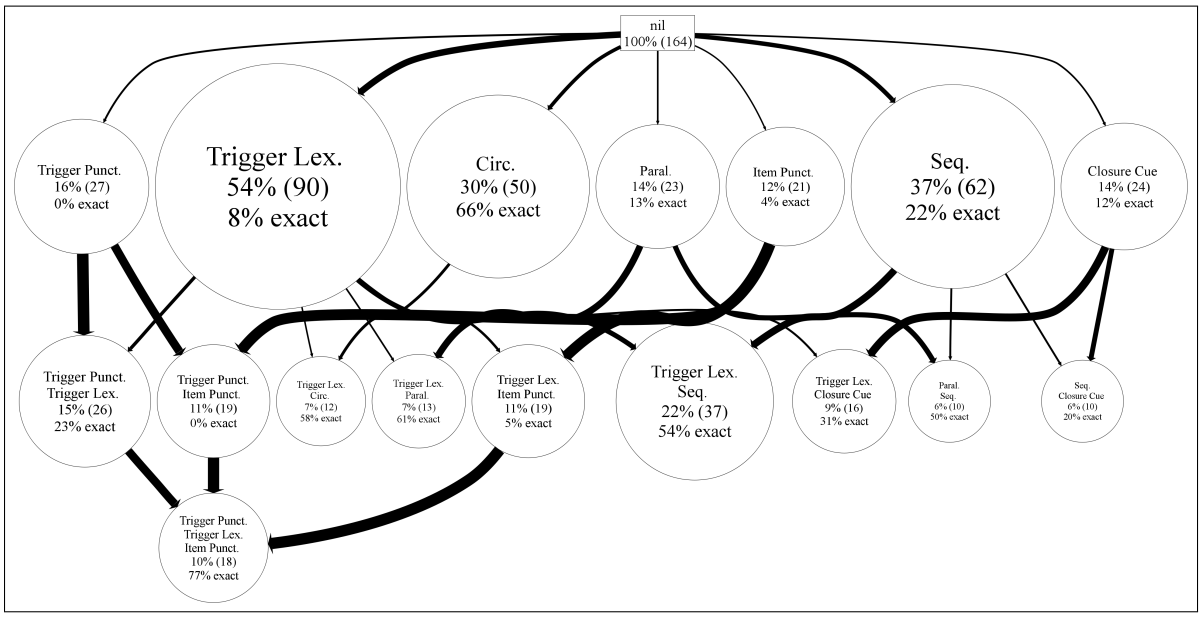

FIGURE 5.4 – Treillis des types d'indices associés aux structures énumératives
Les nœuds correspondent à des cooccurrences d'indices, du plus générique au plus spécifique
Le nombre et la proportion de SE sont indiquées pour chaque configuration
Le second pourcentage indique le nombre de SE qui n'ont que ces types indices
Trigger Punct/Lex : indice ponctuationnel ou lexical dans l'amorce
Item Punct : indice ponctuationnel dans les items
Seq. : marqueur d'intégration linéaire
Circ : circonstanciel
Paral. : parallélisme syntaxique entre les items
Closure Cue : indice lexical de clôture

semblent auto-suffisants (66% des structures qui ont ce type d'indice s'en contentent), les séquenceurs exigent un autre indice, principalement un indice lexical dans l'amorce (seules 22% n'ont que ce type d'indice).

– La combinaison la plus complexe qui soit suffisamment fréquente pour avoir été repérée (les configurations représentant moins de 10%des cas ne sont pas présentes dans le graphique) est la combinaison située tout en bas : indices ponctuationnels dans l'amorce et les items, et indice lexical dans l'amorce.

Ce travail peut être répété pour les trois autres types de structures énumératives, sur la base de treillis organisés sur le même principe (heureusement générés automatiquement par un programme spécifique).

L'approche présentée ici reste cependant embryonnaire, puisque notamment la complexité initiale des données est bien plus grande que celle qui a été utilisée pour cette première analyse. Pour commencer, nous n'avons considéré que les *types* d'indices différents, et non pas les indices eux-mêmes. Ainsi, nous avons assimilé la présence d'un circonstant dans chaque item d'une structure énumérative au cas où seul un sous-ensemble des items contient un tel indice. De même, la position relative des indices dans la structure n'a pas été prise en

compte, ni les sous-types que l'on peut aisément envisager pour certains d'entre eux (types de circonstants, schéma ponctuationnel particulier, etc.). La complexité et la diversité des configurations aurait alors bien entendu dépassé la capacité de ce type de représentation, ainsi que la quantité de données disponible.

Par contre, l'utilisation de treillis correspond directement à la modélisation qui permettra, à terme, le déploiement de méthodes automatiques de repérage des structures énumératives. Bien que cet objectif dépasse le cadre du projet ANNODIS, il s'agit d'un prolongement facilement envisageable. Sur la base d'indices repérés automatiquement, l'identification de configurations suffisamment spécifiques de ceux-ci, tout en gardant une certaine souplesse dans les contraintes, peut permettre un étiquetage automatique des structures énumératives dans un texte. Dans ce cas il est important de bien sélectionner les nœuds qui permettent un repérage efficace : c'est d'ailleurs avec cet objectif en tête que mon choix s'est porté sur les treillis.

5.1.3 Croiser des données de nature différente : structures énumératives et cohésion lexicale

Pour ne pas quitter le projet Annodis, dont le statut en tant que source de données riches et complexes a été important pour mon travail de ces dernières années, je vais maintenant envisager le cas où un travail exploratoire aborde la question du croisement d'informations de différents types. Dans le cadre du projet Voiladis (Voisinage Lexical pour l'analyse du discours) Cécile Fabre (2010) propose d'étudier les mesures de cohésion lexicale en relation avec la structuration du discours. C'est notamment le sujet de la thèse de Clémentine Adam qu'elle co-encadre avec Philippe Muller de l'IRIT. Les travaux de C. Adam ont porté sur l'élaboration d'une mesure de la cohésion lexicale en utilisant des voisins distributionnels, qu'elle a pour l'instant appliquée à la segmentation thématique, mais qui rencontre des difficultés essentiellement dues à la définition de la tâche de segmentation, et la nécessité de recourir à des modes d'évaluation peu satisfaisants et artificiels comme le découpage d'un texte en sections en lieu et place d'une structuration du discours (Adam *et al.*, 2010).

L'idée du travail préliminaire présenté ici (en collaboration avec Clémentine Adam et Cécile Fabre) est donc d'observer le lien entre la cohésion lexicale et les structures énumératives, considérées ici comme des zones de textes délimitées avec une fonction discursive identifiée et validée. Il s'agit donc d'un point de rencontre entre deux univers très complexes : nous avons vu la richesse des données concernant les structures énumératives, mais la mesure de la cohésion lexicale ne va pas simplifier la tâche. L'idée générale de cette mesure est de calculer dans un texte la densité des liens sémantiques (exprimés à travers une ressource générique) entre des unités lexicales. Il s'agit en fait d'une extension du principe mis en place par Hearst (1997), qui n'exploitait que la notion de répétition des unités lexicales. Le travail de Clémentine Adam consiste à utiliser les voisins distributionnels [4] pour repérer des relations plus complexes et plus étendues.

La représentation classique pour des approches de ce type est une courbe qui indique au fil du texte la densité des relations lexicales dans une fenêtre glissante. Une forte densité indique une zone de texte à forte cohésion lexicale, alors qu'une densité faible indique une rupture (ce sont d'ailleurs ces ruptures qui sont explicitement recherchées dans la segmentation thématique). La première approche a donc consisté en une superposition graphique des

4. Plus précisément la ressource obtenue par analyse !distributionnelle sur l'ensemble des textes extraits de la Wikipedia après analyse syntaxique. *Les voisins de la Wikipedia*, réalisée par Franck Sajous, est disponible sur : `http://redac.univ-tlse2.fr/applications/vdw.html`

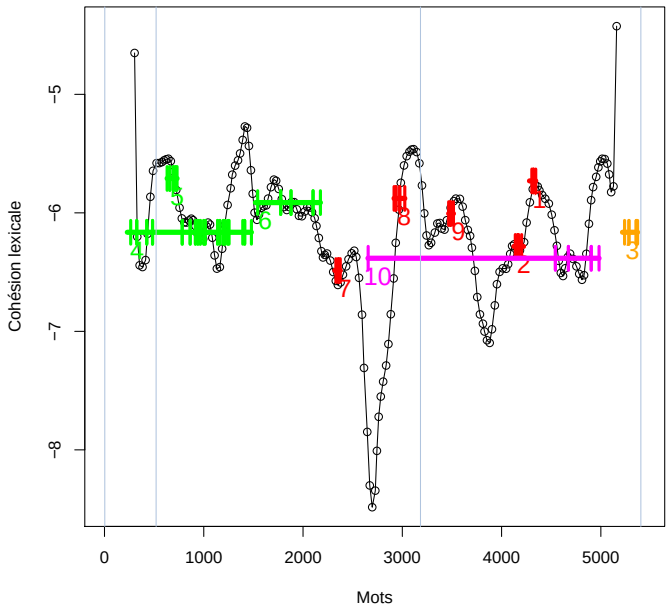

FIGURE 5.5 – Projection des structures énumératives sur un graphe de cohésion lexicale
La courbe mesure pour chaque point du texte la cohésion lexicale, et les structures énumératives sont indiquées par des segments horizontaux numérotés. La couleur des segments indique le type de structure énumérative (1 : magenta, 2 : orange, 3 : vert, 4 : rouge)

deux mondes que nous souhaitions mettre en regard, en indiquant sur la courbe de cohésion les segments de textes correspondant à des structures énumératives. C'est ce qu'indique la figure 5.5.

La courbe de cohésion lexicale est visible en noir, et montre la forme caractéristique de ce genre d'approches sur des textes expositifs longs (ici l'article de la Wikipedia sur Albert Einstein). Les structures énumératives sont indiquées par des traits horizontaux placés en surimpression sur cette courbe en respectant la hauteur moyenne du score de cohésion. Les couleurs identifient les différents types de structures énumératives tels que décrits dans (Ho-Dac *et al.*, 2010). Les deux types d'information sont donc présentés sur une même dimension, celle de la disposition linéaire du texte (les unités de l'axe horizontal sont les nombres de mots à partir du début).

Ce graphique montre que la corrélation, si elle existe, entre les deux phénomènes n'est pas visible : les structures énumératives ne semblent pas apparaître spécialement à des points élevés de la courbe (donc ne correspondent pas spécialement à des zones de forte cohésion lexicale), ni à proximité des « vallées » (donc ne coïncident pas non plus avec les phases de transitions thématiques que ces dernières permettent de repérer). De fait, une première

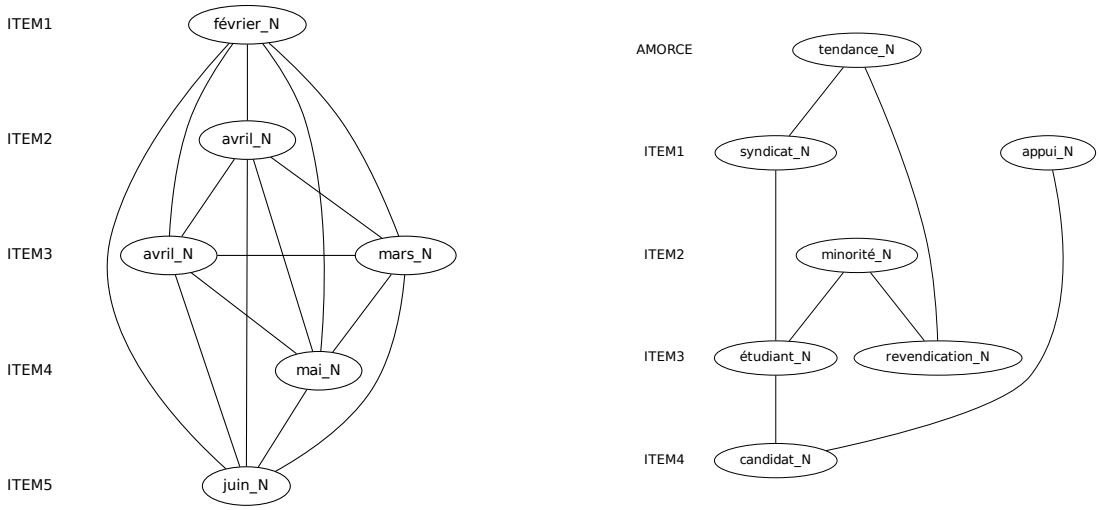

FIGURE 5.6 – Chaînes de relations sémantiques distributionnelles couvrant une structure énumérative

Pour chacune des deux structures, les constituants (amorce, items) sont représentés par les paliers verticaux en respectant l'ordre naturel. Pour chaque pallier sont indiquées les unités lexicales qui ont été repérées et qui entretiennent une relation sémantique distributionnelle avec au moins un autre mot. Ces relations sont représentées par les arêtes du graphe.

expérience pour estimer l'alignement entre les zones de ruptures thématiques et les frontières des structures énumératives n'a pas été concluante.

Une deuxième approche s'est concentrée cette fois sur l'étude qualitative des relations lexicales repérables dans une structure énumérative. Étant donné le nombre et la variété des relations lexicales identifiables par une analyse distributionnelle (ce qui constitue à la fois leur richesse et leur difficulté d'utilisation), la représentation de leurs instanciations en corpus est très difficile (voir les exemples d'Adam et Morlane-Hondère (2009)). Si l'on ajoute à cela la représentation visuelle des structures énumératives, constituées de plusieurs éléments typés et de taille très variable, là encore une approche synthétique est nécessaire. Notre choix a été de définir des contraintes permettant de filtrer les informations disponibles, et de rechercher explicitement des configurations denses entre les relations sémantiques et la couverture d'une structure énumérative. Plus précisément, nous avons extrait les intersections telles qu'une chaîne connexe de relations distributionnelles recouvre toutes les composantes d'une structure énumérative. On garantit ainsi la découverte de composantes thématiques associées à l'intégralité d'une énumération, puisqu'elles doivent couvrir toute son extension. Plusieurs types de configurations sont issues de ce premier filtrage des données, nous en présentons ici deux exemples en figure 5.6. Les mots de la SE qui entretiennent des relations de voisinage distributionnel sont identifiés, et positionnés dans le graphe en fonction de l'élément de le SE dans lequel ils apparaissent, et les arêtes qui les relient indiquent cette relation de voisinage.

La figure de gauche fait ressortir une classe lexicale très dense (une clique, chaque unité lexicale étant voisine de toutes les autres) correspondant aux mois de l'année. La structure

énumérative correspondante présente effectivement une organisation temporelle, décrivant une partie de la chronologie de l'affaire du WaterGate dans la page de la Wikipedia qui y est consacrée (sans autre mode d'organisation globale). Celle de droite montre une organisation plus thématique, avec une structure énumérative issue du même document dont voici le texte et la structure :

> *[En 1968, quatre* ***tendances*** *s'affrontent lors des primaires démocrates]*$_{AMORCE}$*.* *[Hubert Humphrey a l'****appui*** *des* ***syndicats*** *et de l'appareil du parti ;]*$_{ITEM1}$ *[Robert Kennedy séduit les* ***minorités*** *noire et catholique ;]*$_{ITEM2}$ *[Eugene McCarthy porte les* ***revendications*** *des* ***étudiants*** *et des pacifistes ;]*$_{ITEM3}$ *[et enfin George Wallace, ségrégationniste du Sud, opposé aux Droits civiques, se présente comme* ***candidat*** *indépendant.]*$_{ITEM4}$

Si les relations lexicales sont plus ténues, à la fois en termes de structure observée (graphe peu dense) et de relations sémantiques, la structuration thématique autour des groupes politiques transparaît. Bien entendu, de nombreuses autres unités lexicales relevant du même domaine pourraient être ajoutées à celles qui ont été repérées (*primaire*, *démocrate*, *pacifiste*, etc.), mais elles n'ont pas été identifiées par l'analyse distributionnelle.

Cette façon d'examiner les données ouvre bien entendu des pistes à explorer plus systématiquement, mais montre bien la nécessité de se doter de moyens d'observations spécifiques afin de cibler le fonctionnement complexe entre ces deux types de données. Il est notamment envisageable de caractériser automatiquement les structures énumératives sur cette base, à la fois en termes de niveau de cohésion des structures (en fonction du nombre de sous-graphes ainsi repérés et de leur couverture), mais aussi pour en identifier le mode d'organisation.

5.2 Visualiser la disposition dans les textes

Le second type d'approche que je vais détailler ici a déjà été évoqué dans les parties précédentes. Il concerne l'intérêt que présente l'observation de la distribution au fil d'un texte de différents phénomènes langagiers repérables automatiquement. Il peut s'agir de la simple observation d'unités lexicales, de structures discursives, ou encore de la mesure dynamique de caractéristiques textuelles plus globales.

L'intérêt de ce type d'approche est triple :

- il permet de mieux observer le fonctionnement du phénomène ainsi visualisé sous la forme d'une courbe, et notamment de comparer les distributions de phénomènes différents au sein d'un même texte ;
- il permet de caractériser le texte d'une façon globale en observant les caractéristiques distributionnelles ;
- il peut à terme permettre de (ou du moins donner des pistes pour) mettre en place des procédures automatisées pour l'exploitation des textes.

5.2.1 Classes sémantiques

La notion d'isotopie telle qu'elle est utilisée en sémantique différentielle nous vient de Greimas (1966) mais a surtout été promue par Rastier (1987). Définie comme la récurrence d'un trait sémantique (sème), elle peut être vue comme la projection sur l'axe syntagmatique de relations paradigmatiques entre des unités signifiantes. Bien qu'elle soit définie dans la sémantique des textes essentiellement comme un élément du processus interprétatif, au sens

où c'est l'hypothèse de cette récurrence qui permet d'instancier des sèmes, elle peut être objectivée comme une trace obtenue en cours de parcours, et concrétisée dans le texte lui-même comme un ensemble d'occurrences reliées par une similarité sémantique. Ainsi ramenée à des considérations plus pratiques, l'isotopie prend dès lors une forme visible lorsque l'on projette sur les unités d'un texte des classes d'équivalence. On peut ainsi la représenter par une courbe ou une série de points répartis sur l'axe du texte, comme je l'avais proposé pendant ma thèse (voir figure 1.1, page 31).

Un autre exemple de représentation est celui que j'avais utilisé comme dispositif expérimental pour vérifier globalement des traductions dans le cadre du projet Idol (voir section 1.3.3, page 36, et aussi (Tanguy *et al.*, 1999b)). La figure 5.7 montre la disposition relative de quelques isotopies dans un bitexte (i.e. un texte et sa traduction). Pour chaque classe sémantique, les termes de chaque langue ont été rassemblés (ici manuellement à partir d'un lexique de transfert) et projetés sur chaque composante du bitexte. Les zones dans lesquelles des termes d'une classe ont été trouvés sont visibles par une montée de la courbe (avec un seuil inférieur de distance entre les occurrences pour considérer qu'il y a contiguïté).

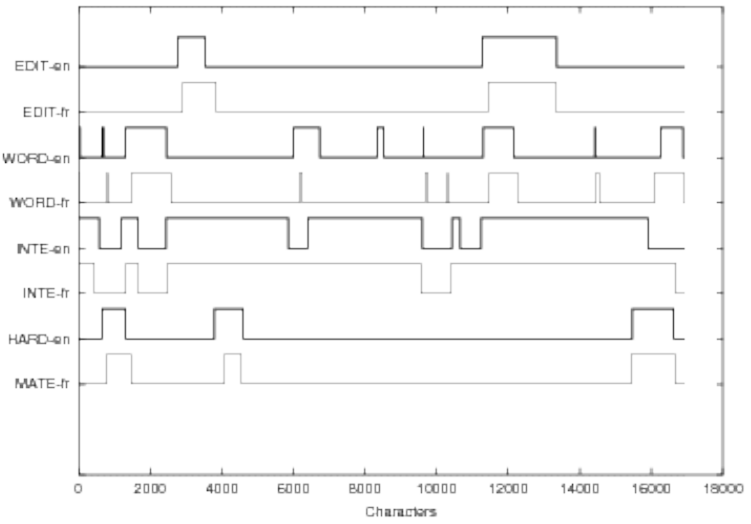

FIGURE 5.7 – Isotopies dans un bitexte
Pour chaque classe sémantique, une paire de courbes indique la disposition des termes correspondant dans chaque texte

Dans le cadre de la vérification de traduction, il s'agissait donc de comparer chaque courbe avec sa correspondante dans l'autre langue pour voir si les dispositions étaient similaires. Une différence pouvait en effet être due à un problème plus global de traduction (déplacement de segments de texte, ou oubli d'un passage).

Hormis ce cas assez précis, on peut identifier d'autres types d'interprétation basés sur l'analyse visuelle des dispositions de classes sémantiques sur l'axe d'un texte :

– Distinction entre classes locales et globales : en fonction de la zone du texte concernée, on peut identifier des situations où la classe sémantique n'a qu'une présence locale, ou au contraire est répartie sur l'ensemble du texte. En analyse du contenu, cela peut permettre d'identifier les thèmes principaux d'un texte (second cas) ou les principaux segments de celui-ci (premier cas). En fonction du mode de représentation choisi, on peut

également étudier la densité des zones du texte en fonction du nombre d'occurrences concernées par la classe sémantique.

– Dispositions relatives dans un texte : en prenant en considération deux classes et non pas une seule, on peut envisager les rapports topologiques qu'elles entretiennent. C'est cette fonction que j'avais essentiellement visée dans ma thèse, pour caractériser deux classes sémantiques dans un même texte, en identifiant des relations particulières : les deux isotopies peuvent être entremêlées (elles recouvrent peu ou prou la même zone du texte), juxtaposées (la zone de l'une est située juste après celle de l'autre) ou totalement disjointes. Sur un plan plus local, une alternance régulière peut traduire la notion de rythme sémantique, prisée par les adeptes de la sémantique textuelle appliquée aux textes poétiques (voir par exemple Missire (2005)).

5.2.2 Structures discursives

On a vu précédemment que la projection des structures énumératives du projet Annodis dans un texte constituait la première façon de les croiser avec d'autres informations. Dans l'expérience que nous allons présenter ici, il s'agit cette fois d'investiguer les relations qu'entretiennent ces structures entre elles.

Le projet Annodis, ou plus précisément la partie de ce projet qui aborde les structures discursives macro-textuelles, a produit un corpus annoté qui contient, en plus des structures énumératives (SE) déjà présentées, des structures appelées Structures à Unité Référentielle (ou SUR). Ces structures plus simples que les premières correspondent aux chaînes topicales qui ont été identifiées par les annotateurs comme séries d'unités ayant le même référent, et qui coïncident avec des segments de texte dans lesquels ce référent est le point focal du discours. Ce référent peut être de différentes natures, mais il se traduit obligatoirement par plusieurs expressions, dont des reprises anaphoriques. Nous ne nous intéresserons ici qu'à la zone couverte par ces structures, et non pas à ces marqueurs.

L'objectif de ce projet était de disposer d'un corpus pour observer à large échelle le fonctionnement de ce type de structures : on a vu que pour les structures énumératives cela a permis l'établissement d'une typologie et l'identification d'un ensemble de caractéristiques (taille, composition, types d'indices, etc. voir le détail en section 6.2). Ces structures étaient initialement peu spécifiées et leur définition se basait sur des hypothèses nécessitant une validation empirique. Notamment, il était attendu que les structures énumératives se trouvent à différents niveaux de granularité (de la section complète jusqu'à l'intra-paragraphique) et qu'elles pouvaient être enchâssées. Par contre, le niveau et la fréquence de ces enchâssements restait à mesurer, c'est ce que nous avons fait en utilisant une visualisation à l'échelle du texte.

La figure 5.8 montre ainsi les zones couvertes par les structures énumératives (en haut) et les structures à unité référentielle (en bas), le long de l'axe du texte. Lorsqu'il y a enchâssement, c'est-à-dire lorsqu'une structure débute à l'intérieur d'une autre, elle est positionnée à un niveau supérieur sur le graphique : le nombre d'étages ainsi obtenus permet de mesurer facilement le niveau d'enchâssement d'un segment de texte.

Le rectangle attire l'attention sur le niveau impressionnant de structures énumératives enchâssées dans ce texte particulier : il s'agit d'un article scientifique écrit par mon collègue Michel Roché (Roché, 2008) qui établit le record avec une profondeur de 5 (il s'agit pour sa défense de la présentation d'une typologie à plusieurs niveaux). En échange, les structures à unité référentielle sont petites et sporadiques, comme l'indique la partie basse de la figure.

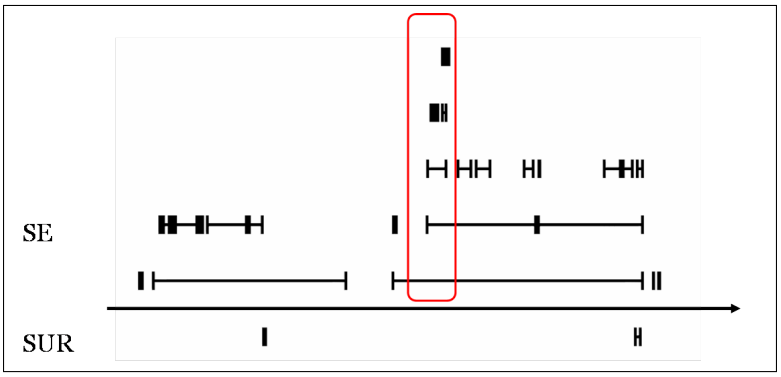

FIGURE 5.8 – Positions relatives des structures énumératives (SE) et des structures à unité référentielle (SUR) dans un texte (premier exemple).
Chaque structure du texte est représentée par un segment horizontal correspondant à la zone de texte couverte ; la disposition verticale traduit l'enchâssement des structures.

De même, il semblerait qu'il n'y ait pas d'interaction particulière entre les deux types de structures.

Chaque texte peut ainsi être rapidement et facilement étudié sur cet aspect particulier, et la comparaison entre les textes peut se faire sur cette base, tout comme elle peut permettre à l'échelle du corpus une observation de la variété des configurations. Ainsi, la figure 5.9 montre un autre cas où, cette fois, aucun enchâssement n'est visible mais où, par contre, les structures à unité référentielle sont bien plus fréquentes, et semblent alterner avec les structures énumératives comme mode d'organisation discursive.

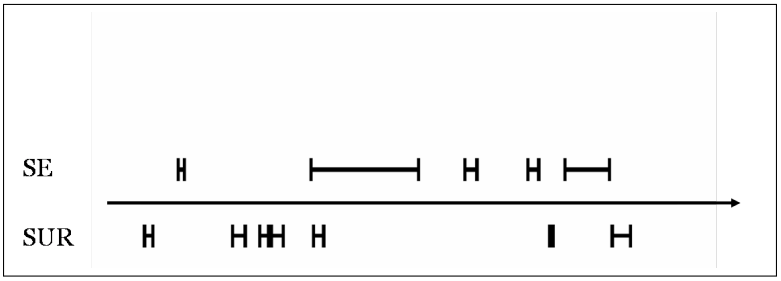

FIGURE 5.9 – Positions relatives des structures énumératives (SE) et des structures à unité référentielle (SUR) dans un texte (second exemple)

Ces différentes configurations permettent d'identifier des cas particuliers pour certains textes comme on l'a vu dans le premier cas, mais aussi à observer des régularités à travers les textes du corpus, ou encore à croiser ces informations avec des types de textes ou des types de structures discursives. Ce type de représentation est encore une fois un allié précieux lorsque l'on souhaite, comme c'est le cas ici, disséquer au plus près un type d'objet textuel.

5.2.3 Appels de citation dans les articles scientifiques

Dans le cadre du projet Rhecitas (Rhétorique des relations de citations dans les articles scientifiques), je me suis intéressé aux contextes des appels de citation bibliographique dans les publications scientifiques, et plus particulièrement dans celles des sciences humaines et sociales.

L'analyse des citations est un champ d'investigation important dans les sciences de l'information : on sait l'importance croissante qu'ont pris les indices bibliométriques dans les processus d'évaluation de la recherche scientifique à l'échelle du chercheur, du laboratoire ou du support de publication. Toutefois, ces mesures s'effectuent sur des données extrêmement pauvres, et se contentent d'exploiter les listes de références sans s'intéresser aux contextes dans lesquels ces dernières sont insérées dans le texte de l'article. C'est principalement les travaux de Simone Teufel et de ses collègues (Teufel *et al.*, 2006) qui ont inspiré ce projet. Notre approche reprenait les principes de leurs travaux, en mettant en place une étude contextuelle des appels de citation destinée à identifier les marques linguistiques qui permettraient de discriminer entre les différentes fonctions rhétoriques de celles-ci (voir (Tanguy *et al.*, 2009) pour plus de détails sur ces aspects, et également la section 7.2.3). De telles fonctions comprennent par exemple : la référence à un cadre théorique, l'utilisation de données ou d'outils empruntés à un tiers, la comparaison avec une autre méthode, etc.

Nos ambitions ont toutefois dû être revues à la baisse au vu de la difficulté d'une classification manuelle pour les publications de sciences humaines : si tous les travaux précédents de ce type se sont penchés sur des publications en sciences « dures » (S. Teufel elle-même ayant traité les publications en TAL et en chimie) ils avaient pu bénéficier de la structure prototypique voire normée de celles-ci, bien souvent rendue nécessaire par les politiques éditoriales des journaux et des conférences. En SHS, les types de contexte, les rôles attribués aux citations elles-mêmes, et même la structure générale d'un article sont extrêmement variables, et font parfois intervenir des informations implicites très complexes à démêler, même avec une lecture attentive. Bref, la tâche ne pouvait simplement pas être abordée dans ce cas, et nous avons décidé de la simplifier à l'extrême en ne cherchant à distinguer que les citations de premier plan (nécessaires à la compréhension de l'article) à celles d'arrière-plan (remplissant des fonctions secondaires).

Malgré la modestie de la tâche, cette opération a demandé un gros effort sur le traitement des données, le plus complexe étant le repérage des appels de citation dans le texte, procédure qui doit être d'une robustesse à toute épreuve pour s'adapter aux différentes normes (et absences de norme) utilisées. Là encore, l'emploi systématique d'outils de gestion de référence (notamment Bibtex) par les auteurs de publications de sciences dures avait masqué à nos prédécesseurs la difficulté énorme de cette tâche dans les publications de SHS, où il semblerait bien que la majorité gère ses références bibliographiques à la main, ce qui entraîne inévitablement des erreurs et des incohérences dans leur notation. Quoiqu'il en soit, le travail effectué par nos collègues de l'INIST a permis d'avoir des résultats parfaitement exploitables pour quelques centaines d'articles extraits du portail `revues.org`.

Il semble qu'avant même de regarder en détail le contexte linguistique d'insertion des appels de citations, et notamment en y recherchant les différentes marques du positionnement de l'auteur (ce que nous avons également fait, voir partie IV), la simple position de ces appels dans le texte peut être un critère déterminant. C'est d'ailleurs ce qu'avait repéré S. Teufel, en utilisant la position relative comme trait utilisé dans son approche par apprentissage automatique (mais là encore, la structure imposée des articles scientifiques, qui exigent notamment en

TAL une partie « travaux similaires » à la fin de l'article facilitait la tâche). Par contre, mon intuition est que l'étude de la répartition dans le texte de l'ensemble des appels de citation (quelle que soit la référence exacte) peut être une indication relative à la nature de l'article.

La figure 5.10 montre ainsi quelques exemples extraits du corpus du projet Rhecitas.

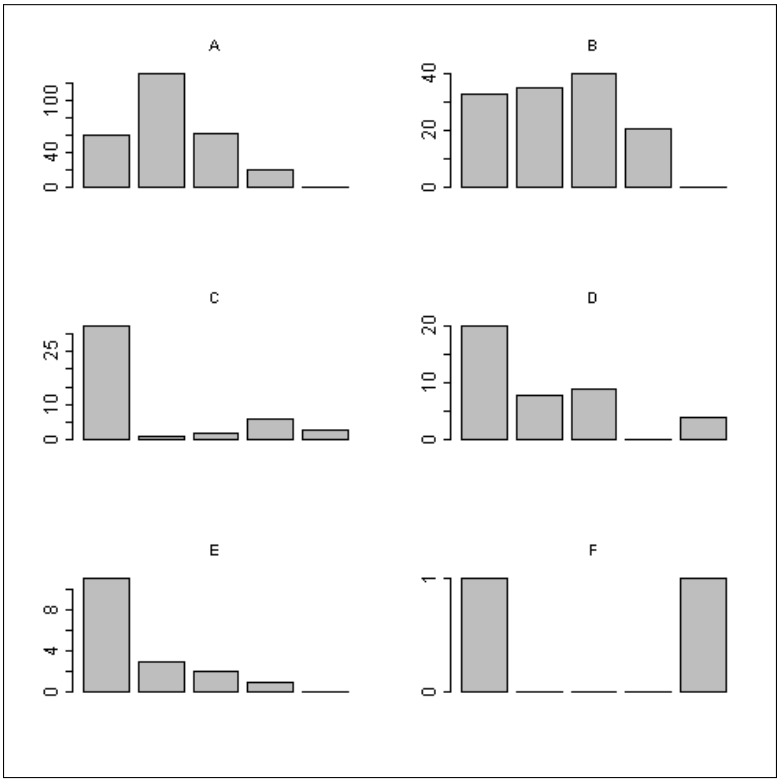

FIGURE 5.10 – Différents schémas de répartition des appels de citation dans un article scientifique

Pour chacun des 6 articles présentés, découpés en 5 parties égales, les barres indiquent le nombre d'appels de citation présents.

Chaque texte a simplement été découpé en 5 parties égales, et la fréquence des appels de citations a été mesurée dans ces 5 parties. Tant en termes de niveau (attention, l'échelle de l'axe vertical varie d'un cas à l'autre) que de répartition, on voit d'ores et déjà que plusieurs profils graphiques peuvent être dégagés. Les 6 articles choisis correspondent en effet à des catégories très différentes :

- les textes A et B sont des états de l'art : le volume global des appels de citations est très important, et ils sont répartis uniformément dans l'ensemble du texte à l'exception de la partie finale ;
- le texte C est la présentation d'un modèle théorique par son auteur, donc avec peu de références sauf en tout début d'article (pour expliciter son positionnement théorique) ;
- les texte D et E sont des cas classiques de description d'une expérimentation, avec un cadrage théorique initial, puis des citations permettant de décrire les étapes, et enfin les interprétations qui font peu appel à des travaux cités ;
- le texte F est un cas très spécifique et marginal d'un récit de vie, avec une absence

quasi-totale de références.

Bien entendu il existe de nombreux facteurs à prendre en compte pour de telles caractérisations, mais il me semble que ce type de considérations peut facilement être intégré dans les approches d'analyse des productions scientifiques qui commencent à se généraliser tant pour des applications en sciences de l'information et de la documentation que pour des études linguistiques, comme par exemple dans le projet Scientext (Tutin *et al.*, 2009).

5.2.4 Interaction dans les consultations médicales

Le dernier exemple de représentation synthétique que j'ai pu proposer dans un cadre spécifique concerne à nouveau le projet Intermede, mais cette fois pour traiter les données correspondant à l'analyse de notre propre équipe. Comme je l'ai déjà dit dans la première partie, dans ce cadre interdisciplinaire les aspects liés à l'analyse lexicale et thématique étaient traités par des collègues sociologues et psychologues qui ont utilisé des outils d'analyse de contenu (ModaLisa, Alceste et Lexico). Nous, linguistes de l'ERSS [5], avons alors proposé de nous intéresser à une dimension linguistique qui n'est pas traitée (ni sans doute traitable de façon simple ou satisfaisante) par de tels outils : l'interaction. Les objectifs de ce projet concernaient la mise au jour de variations dans les interactions médecin/patient dont on pourrait tester la corrélation avec des caractéristiques non-linguistiques (âge, sexe, niveau social, etc.). Nous avons abordé cette question par le biais de mesures spécifiques et d'analyses statistiques qui seront présentées au chapitre suivant, mais nous avons aussi proposé un mode de représentation particulier pour aider notre analyse (Tanguy *et al.*, 2011a; Vergely *et al.*, 2009).

Si les dialogues et les échanges sont un objet d'étude très exploité en linguistique, nous n'avons pas trouvé de travaux précédents qui abordent ces données sur le plan de l'évolution d'une interaction langagière. Les principaux efforts concernent en effet la modélisation des phénomènes locaux de communication (reformulation, interruptions, questions/réponses, etc.). Nous avons abordé ces questions nouvelles pour nous sur le plan d'une première approche quantifiée que je vais présenter ici.

La première distinction qui nous est apparue importante (et qui a été confirmée par des analyses statistiques) est la variation de répartition de la parole entre le patient et le médecin. Si la situation particulière de communication est basée sur une inégalité de fait (de savoir, d'autorité) entre les deux protagonistes, dans la pratique des types de consultations très différents ont été identifiés dans les analyses du corpus. Entre le médecin très dirigiste qui monopolise la parole, et le patient qui vient chercher plus une écoute qu'une réponse médicale, des situations très contrastées peuvent facilement se dégager d'un simple calcul de la proportion du nombre de mots prononcés par chaque locuteur.

Mais il est également très éclairant de regarder comment ce rapport évolue au fil de la consultation. J'ai donc proposé une représentation graphique permettant de visualiser cette évolution en utilisant une technique de fenêtre glissante dans le texte balisé de la consultation. A différents points du texte (tous les 50 mots), la proportion de ceux prononcés par le médecin a été calculée. La courbe a ensuite été lissée linéairement, et le résultat est visible dans la figure 5.11 pour la consultation du patient Eric. À cette courbe principale (en trait continu) j'ai ajouté :

5. Ce travail a été réalisés en collaboration avec Anne Condamines, Cécile Fabre, Amélie Josselin-Leray et Pascale Vergely.

- les numéros des tours de parole sur l'axe des abscisses, ce qui permet un retour plus facile vers le texte ;
- les occurrences des questions posées par le médecin et le patient (en haut) ;
- le nombre moyen de changements de locuteur et donc l'inverse de la taille des énoncés (courbe en pointillés) ;
- le découpage en phases de la consultation par analyse manuelle de l'ensemble des transcriptions. Ce découpage reprend les phases typiques d'une consultation de médecine générale identifiée par Have (1989).

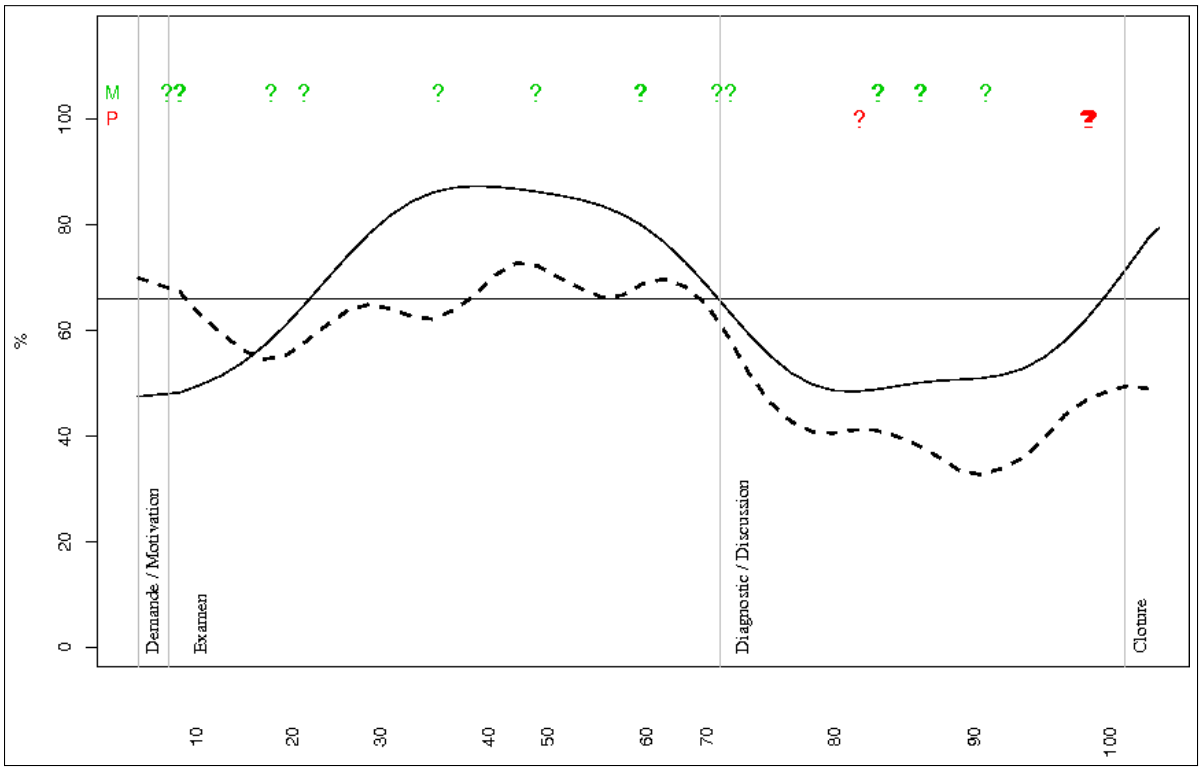

FIGURE 5.11 – Profil de consultation médicale (Eric)
Sont présentés le long de l'axe du texte : la proportion de mots prononcés par le médecin (trait continu) ; le nombre de changements de locuteur (trait pointillé) ; les questions posées par chacun des deux locuteurs (« ? ») ; les phases de la consultation (traits verticaux).

Le cas de la consultation d'Eric est très typique d'une consultation « normale ». Comme on le voit, la répartition de la parole suit exactement le découpage en phases :

- dans la phase initiale (très courte ici) c'est le patient qui parle le plus puisqu'il expose son problème ;
- la phase d'examen voit le médecin prendre progressivement le monopole de la parole, avec un grand nombre de questions ;
- la phase de discussion, où le médecin expose son diagnostic et explique le traitement est une phase plus interactive, le partage de la parole revient à l'équilibre, et le patient pose à son tour des questions.

Dans d'autres cas, cet aspect prototypique est très loin d'être visible. La consultation de

Mary (figure 5.12) montre en effet un profil totalement différent, bien plus accidenté. Les phases n'ont d'ailleurs pas pu y être repérées, puisque la consultation ne suit absolument pas le déroulement classique (la patiente vient d'accoucher : elle raconte la naissance et sa vie actuelle à son médecin généraliste). On peut y repérer vers le tour de parole n°265 une phase de monologue importante de la part de la patiente (parole du médecin basse, tours de parole longs), lorsqu'elle raconte un accident survenu à un de ses enfants.

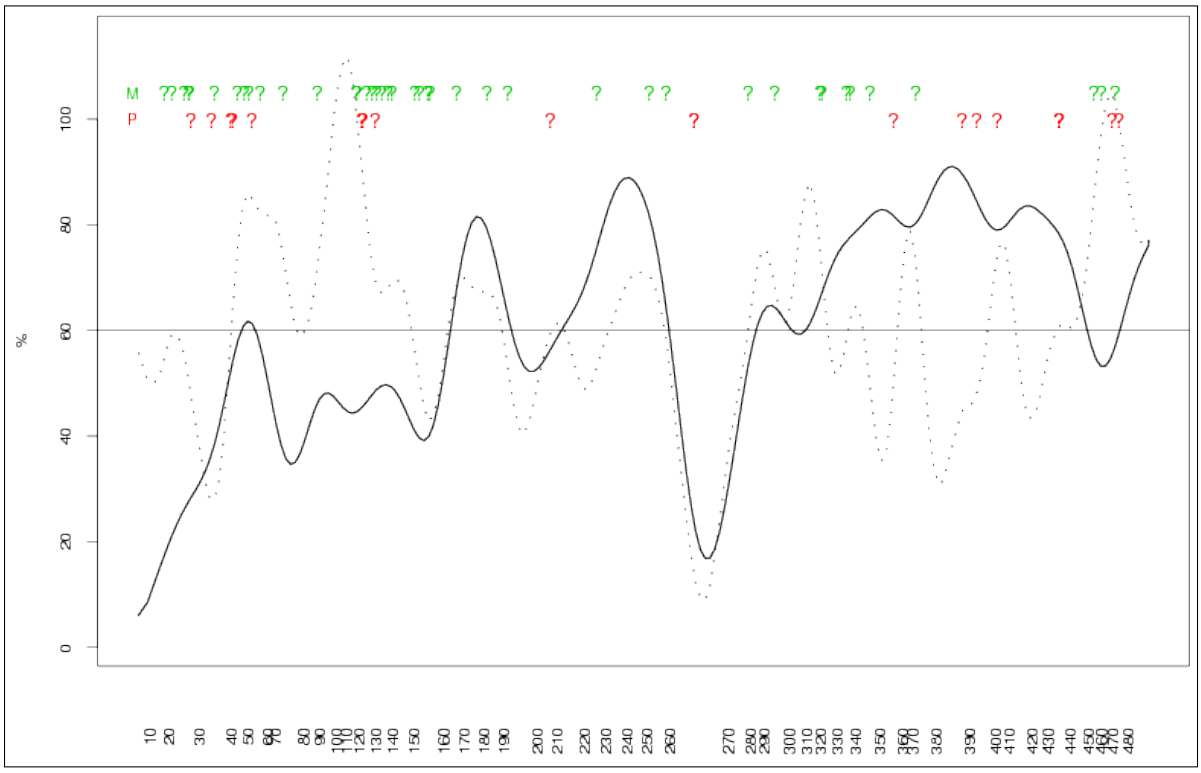

FIGURE 5.12 – Profil de consultation médicale (Mary)

Comme on le voit, ce type de représentation permet donc une caractérisation globale rapide des consultations, mais elle permet aussi d'y identifier des lieux particuliers. Dans ce deuxième cas, le retour au texte est bien entendu nécessaire, et une fonctionnalité souhaitable serait la possibilité d'y naviguer à partir de la courbe. Bien entendu, cette opération est impossible sans mettre en place un outil dédié, ce qui représenterait un coût totalement disproportionné pour une méthode expérimentale comme celle-ci.

Une autre possibilité offerte par cette première expérience est celle d'une segmentation automatique des consultations, sur le principe de la segmentation thématique (qui se base sur le repérage des variations importantes dans la mesure de cohésion lexicale). Ici, les phases prototypiques pourraient être repérées par un calcul sur les différentes caractéristiques évoquées sur la courbe, une fois les consultations trop atypiques écartées.

Ce corpus très riche a également donné lieu à la mesure d'autres phénomènes liés à l'interaction (répétitions, reprises, énoncés de continuation, etc.), mais leur analyse a mobilisé d'autres techniques plus traditionnelles qui seront présentées dans le prochain chapitre.

On a ainsi pu voir à travers ces quelques exemples l'intérêt des visualisations pour répondre à différentes questions sur les données langagières. J'ai souvent dû improviser et tâtonner, et obtenu des résultats qui n'étaient pas toujours entièrement satisfaisants. Il me paraît donc important à ce stade de prendre du recul par rapport à ce type de travail, et surtout de réfléchir à la nature de cette activité et à son articulation avec d'autres modes d'exploration des données.

5.3 La visualisation des données : un enjeu majeur pour la linguistique

Il existe une petite tradition de représentations graphiques des données linguistiques, et celles-ci ont été multipliées par le développement du TAL : les arbres syntaxiques, les graphes de mots pour représenter les relations lexicales, les spectrogrammes des signaux de parole, les hiérarchies de concepts et de terminologie, etc. Widdows (2004). Ces différents types de représentation sont utilisés à la fois comme tels pour leur interprétation facilitée par les linguistes, comme outils de modélisation, ou encore comme métaphores pour mettre en place des paradigmes (comme c'est le cas des modèles vectoriels de représentation des documents).

Mais dans ce type d'activités également, la complexification et la massification des données ont créé de nouveaux besoins en termes de techniques de visualisation. Il est donc important à mes yeux de voir comment on peut rapprocher nos considérations disciplinaires des questions que se posent également d'autres domaines de la science et de leurs applications. Le champ de la visualisation des données est en effet en pleine ébullition, prenant appui sur les capacités techniques des logiciels qui visent à assister l'analyse de données massives et complexes par la mise en place de représentations graphiques. En me basant sur quelques travaux et réflexions issues de ces travaux interdisciplinaires, je vais tâcher de résumer quelques points vitaux pour la linguistique empirique.

5.3.1 Besoins en visualisation

On a pu voir dans les exemples présentés dans ce chapitre les différentes raisons qui m'ont poussé à avoir recours à une visualisation des données, je vais les récapituler ici, et en indiquer d'autres qui, si je ne les ai pas spécifiquement détaillées, sont bien présentes dans un grand nombre de situations.

La première question est bien entendu celle de la masse d'information à observer, ce qui rejoint les questions sur l'interrogation de grands corpus abordées dans la partie précédente. Le but recherché par la visualisation est donc une vue d'ensemble de ces données, qui doivent être appréhendées globalement, et pas simplement sélectionnées, listées ou comptées. Cette notion de « vue aérienne » est par exemple celle qui est utilisée pour visualiser des données hiérarchiques dans le dispositif des *treemaps* (Johnson et Shneiderman, 1991), en représentant sur une surface bidimensionnelle des unités imbriquées, comme les répertoires et les fichiers stockés sur un disque dur. Tout en permettant l'accès à des informations ponctuelles, ce type de représentation a surtout comme avantage de permettre une lecture globale, et par exemple de repérer rapidement les grandes unités englobantes et leur structure interne.

Un point important soulevé par cet exemple est le fait que les données à visualiser entretiennent très souvent des relations entre elles, que la représentation doit bien entendu traduire et exploiter. Le cas le plus courant est l'étude des réseaux, et leur représentation sous

la forme d'un graphe. Les applications de ces techniques de visualisation sont légion : l'étude des réseaux sociaux, des documents hypertextuels, des bases de données bibliographiques, des dictionnaires de synonymes, etc. La valeur ajoutée de ces représentations concerne l'identification de configurations globales d'une collection. Les principales caractéristiques aisément repérables sont les sous-ensembles denses (zones du graphe isolables du reste, et dont les unités entretiennent entre elles un grand nombre de relations) et les individus-pivots (qui se situent à la frontière de plusieurs des zones denses). L'identification de ces lieux particuliers d'un graphe est rendue possible par l'utilisation d'algorithmes de projection planaire d'un graphe qui cherchent spécifiquement à minimiser le nombre de croisement de liens, et privilégieront donc un placement resserré de ces groupes denses. On a vu ce type d'interprétation dans la section 5.1.1.2. Toutefois, et en dépit des dispositifs techniques déployés, le passage à l'échelle de la représentation d'un graphe comportant des milliers ou plus de nœuds pose des problèmes de lisibilité, comme l'indiquent Henry et Fekete (2008) pour les réseaux sociaux, et nécessite donc d'articuler différents modes visuels.

Un second besoin concerne la multiplicité des caractéristiques attribuées aux données individuelles, qui se traduit par un grand nombre de dimensions à projeter sur une même représentation graphique. Cette projection est nécessaire pour mettre en évidence les liens éventuels entre ces caractéristiques, quelle que soit leur nature. Les possibilités de représentation de dimensions différentes sont bien entendu limitées, mais Bertin (1970) en a identifié huit : les deux dimensions du plan sur lequel est dessinée la représentation, mais aussi la taille, la valeur, la forme, le grain, la couleur et l'orientation des éléments graphiques qui y sont disposés. Le choix principal est alors de définir un espace de référence : si celui-ci est, pour certains domaines, évident car imposé par la nature des données (cartes géographiques, molécules et autres chaînes de l'ADN, corps humain, représentations d'objets physiques comme les véhicules, etc.), il peut également être abstrait ou correspondre à une des caractéristiques non-spatiales. Par exemple, dans les graphes de la figure 5.6, la dimension verticale correspond aux composantes logiques d'une structure énumérative, et dans le treillis de la figure 5.4 il s'agit du nombre de types d'indices. Lors de la représentation planaire d'un graphe, les dimensions ne correspondent à aucune caractéristique des objets particuliers, mais uniquement à leurs interrelations. Les seuls cas où la nature des données impose un choix naturel sont ceux où l'on vise la linéarité du texte comme c'est le cas pour les exemples de la section 5.2.

En ce qui concerne les objectifs de la mise en place d'une visualisation, on peut en distinguer deux grands types.

Le premier correspond à une approche exploratoire, dans laquelle le lecteur n'a pas nécessairement une idée précise des phénomènes à rechercher. La visualisation permet alors de faire émerger des formes « intéressantes » ou particulières, comme des répartitions non homogènes, ou des ruptures dans une courbe.

Dans le second cas, le type de configuration recherchée est connu, mais est considéré comme plus facilement repérable par une visualisation que par un calcul. C'est le cas des nœuds pivots ou des zones denses dans un graphe, ou des points de convergence lorsque l'on superpose plusieurs courbes. Dans la plupart des cas le repérage par des procédures de calcul ciblées est possible mais complexe car il nécessite de fixer un ensemble assez large de paramètres et est source de bruit.

On rejoint bien ici l'intérêt mis en avant par les adeptes de la visualisation des données : cette capacité à provoquer l'intuition du lecteur en s'appuyant sur d'autres facultés cognitives que le calcul. C'est également ce que défend Rastier (1991) en rapprochant les processus interprétatifs (et de façon plus globale, toute approche de la chose sémantique) de la reconnaissance

de formes ; différents arguments en psychologie cognitive appuient ce point de vue.

5.3.2 Nécessité de multiplier les méthodes et les approches

Étant donné la variété des données, des techniques et des objectifs visés, il est rare qu'un seul type de représentation graphique soit suffisant et/ou identifiable comme le meilleur pour une situation donnée. Dès lors, il est important d'envisager d'emblée que l'étude d'un ensemble de données par visualisation va nécessiter de multiplier les approches. Les raisons en sont les suivantes :

- Comme dit précédemment, l'existence d'un référentiel unique et naturel pour une représentation spatialisée ne correspond pas à toutes les situations. Le choix des dimensions principales de la représentation est donc totalement dépendant des objectifs. Et même dans les cas où une entité spatiale tridimensionnelle constitue l'objet d'étude, aucune représentation planaire ne permet d'en traduire la complexité, et il est alors nécessaire, comme en architecture, d'en multiplier les vues.
- Malgré les possibilités de cumul de l'information par différentes modalités graphiques, il est bien entendu illusoire de surcharger une représentation graphique, et celle-ci reste donc toujours une projection limitée de la complexité des données. De ce fait, toute interprétation nécessite d'utiliser différentes combinaisons pour pouvoir espérer les examiner chacune en regard des autres.
- Dans les approches exploratoires surtout, la multiplication des points de vues fait partie intégrante du processus de découverte. Certaines configurations n'apparaissent que sous une certaine combinaison de paramètres de visualisation qu'il est donc important de faire varier.

5.3.3 Intérêts de l'interaction

Comme on le voit, la multiplication des modes de visualisation est nécessaire, non pas simplement pour s'adapter à de nouvelles données, mais bien pour mettre à la disposition de l'analyste différents points de vue sur la même collection d'objets. Il est donc logique que les outils de visualisation de données s'orientent systématiquement vers des dispositifs interactifs, qui donnent à leur utilisateur un ensemble de possibilités permettant de modifier la visualisation. Les libertés ainsi accordées peuvent être de différents types :

- Un choix des dimensions à projeter, et de leur mode de représentation. Il s'agit donc pour l'utilisateur de décider des caractéristiques des objets qu'il souhaite voir représenter, en contrôlant les détails de leur visualisation. On peut voir l'exemple d'une telle interaction dans la figure 5.13, correspondant à l'outil de visualisation FromDaDy de Hurter (2010). Dans le cas qui est présenté, les données correspondent à des positions d'avions, pour lesquelles sont disponibles la longitude, la latitude, l'altitude, la vitesse, le temps et l'identification de l'appareil. Dans la fenêtre de dialogue présentée à gauche, l'utilisateur peut associer ces caractéristiques à des dimensions visuelles, par exemple les coordonnées x et y dans l'espace de présentation, la couleur et la taille des points. Le rendu dynamique est présenté en partie droite de la figure. La figure du bas montre les mêmes données avec une configuration différente, en projetant l'altitude des vols sur l'axe des ordonnées.
  L'utilisateur peut donc à loisir tester différentes configurations de visualisation, lui permettant ainsi de voir apparaître des formes spécifiques interprétables (dans l'exemple

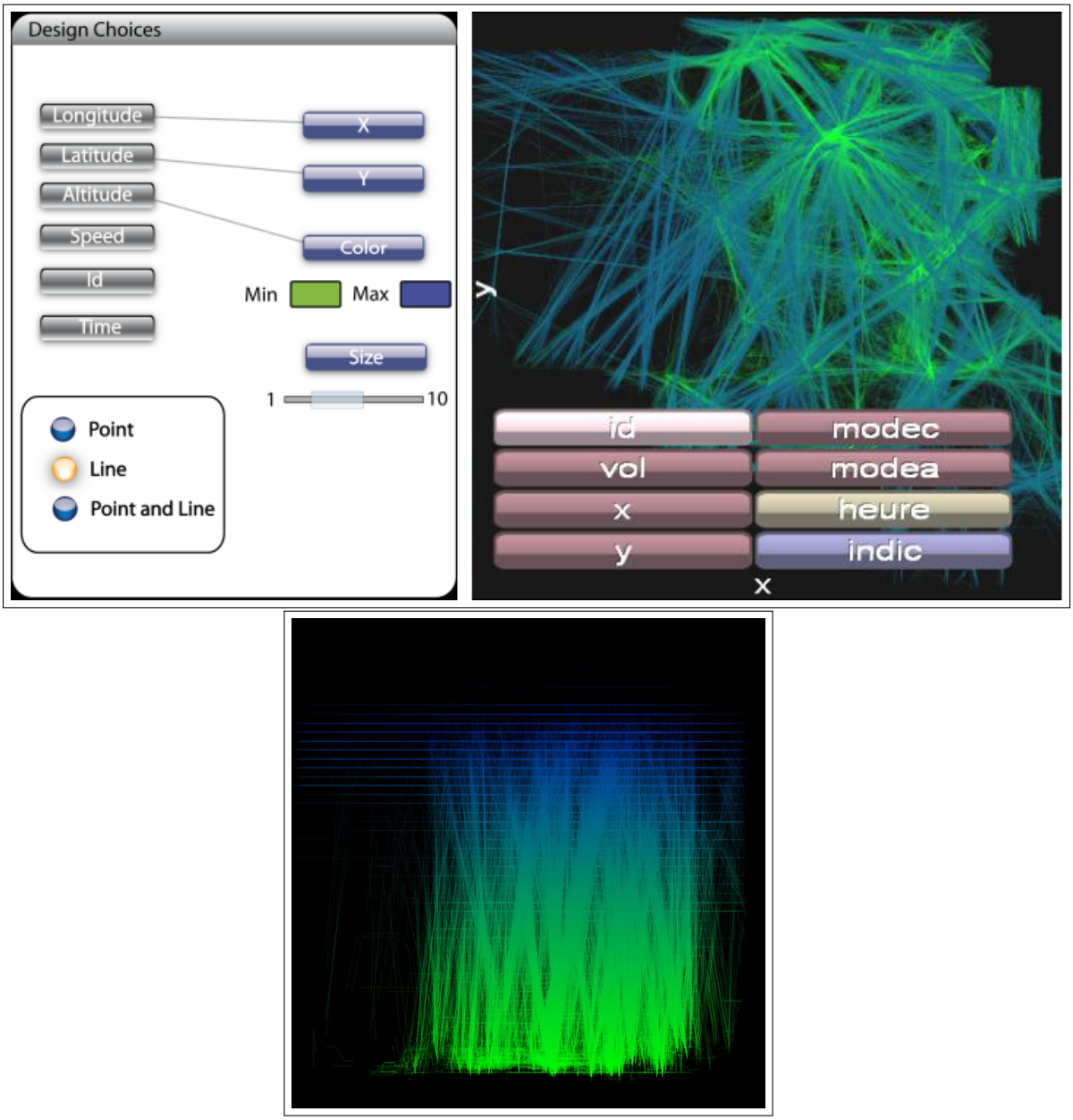

FIGURE 5.13 – Exemple d'outil de visualisation interactif : FromDaDy

de la figure 5.13, des couloirs aériens et des grands aéroports dans la première configuration, et des niveaux de vols pour la seconde). Ce type d'outil est en fait très répandu dans les plate-formes d'analyse des données : la plate-forme Weka par exemple propose un module de visualisation, mais moins sophistiqué (Witten et Frank, 2005)).

– Un ensemble d'opérations de manipulation sur la vue construite. Ces opérations permettent à l'utilisateur de naviguer dans les données à partir de leur représentation visuelle, afin d'affiner son analyse. Ces opérations élémentaires peuvent correspondre à des opérations purement visuelles (zoom, rotation, etc.) ou à des opérations logiques sur les données, dans le même ordre d'idées que celles qui permettent d'explorer une

base de données (sélection d'un sous-ensemble d'individus, réordonnancement, ajout ou retrait d'un sous-ensemble des objets de la collection, etc.). Bien entendu, ces opérations sont d'autant plus complexes qu'elles s'expriment par manipulation directe de l'interface graphique ; la complexité existe autant du point de vue de l'implémentation des fonctionnalités de l'outil que de leur appropriation par l'utilisateur.

– Des réglages sur la visualisation et les données. Même si les tâches les plus importantes semblent être la sélection des dimensions et leur traduction sous une forme visuelle spécifique, il subsiste un ensemble très important de choix locaux à effectuer. Ces choix concernent essentiellement les prétraitements que doivent subir les données brutes pour être associées à une vue particulière : leur mise à l'échelle, leur filtrage, l'établissement d'un seuil, etc. On a vu dans les exemples de graphes dans ce chapitre qu'ils faisaient systématiquement appel à ce genre de réglage, que j'ai moi-même opéré en testant les différents rendus. Ces réglages cherchent généralement à filtrer les données en évitant de surcharger le graphique, tout en préservant le maximum d'information. Là encore, on voit bien que ces choix hautement relatifs et subjectifs sont directement liés à l'interprétation des données visualisées, et sont donc bien mieux effectués par l'utilisateur final que par le concepteur (et dans l'idéal avec un dialogue entre les deux protagonistes pour définir les fonctionnalités).

L'articulation de ces différentes phases de l'analyse et l'ordonnancement des opérations est en phase de devenir une procédure normalisée, pour l'instant généralement exprimée par le mantra de Ben Shneiderman (1996) : *Overview first, zoom and filter, then details-on-demand.*

5.3.4 Limites des méthodes visuelles

Les outils modernes de visualisation des données donnent donc, comme on le voit, une grande place et une grande liberté à leur utilisateur. Ces outils n'en restent pas moins très complexes à manipuler, de par la palette de fonctionnalités qu'ils proposent, et de par les interdépendances des différents choix que l'utilisateur doit effectuer. Ils sont donc clairement destinés pour l'instant à des experts des questions de visualisation plus qu'aux seuls spécialistes des données étudiées.

Si l'on reprend les différentes questions évoquées dans ce chapitre, il apparaît en fait que l'utilisation d'une visualisation des données pour un travail d'investigation requiert une triple compétence :

1. Une connaissance fine et approfondie *des données analysées*. Cette connaissance est nécessaire pour guider les différentes procédures d'investigation, poser les bonnes questions et savoir faire la différence entre des évidences déjà connues (qui ne manquent pas d'apparaître au premier plan de toute représentation) et des faits nouveaux et surprenants.
2. Une expérience des *procédures de visualisation*. En plus d'une familiarité avec les outils utilisés, il est important de disposer de connaissances spécifiques permettant de guider les choix de représentation, par exemple de savoir quelle modalité visuelle associer à une caractéristique particulière pour pouvoir la comparer à une autre. Ce type de compétence semble bien n'être acquis que par une confrontation intensive avec des données et des modes de représentations variés.
3. Une compétence en *traitement des données*. Il ne faut, là encore, ni ignorer ni sous-évaluer les besoins plus techniques qui permettent au final d'associer un outil de visualisation avec des données spécifiques. Cela concerne un ensemble de considérations, allant

des transformations de formats de données aux procédures de normalisation et de traitement statistique des valeurs étudiées. Comme on l'a vu dans de nombreux exemples dans ce chapitre, les caractéristiques des données proviennent très souvent d'un calcul préalable, dont les paramètres ne sont pas du ressort de l'outil de visualisation, mais doivent être réglés en amont (et réglés à nouveau en s'appuyant sur le retour précieux qu'offre là aussi la visualisation).

Si tous ces points peuvent être résolus par une accumulation d'expériences, et/ou en collaborant avec des spécialistes des différentes questions (et les spécialistes de visualisation de données semblent faire preuve d'enthousiasme quand il s'agit de s'attaquer à une situation nouvelle), le dernier problème est plus difficile à régler.

Il concerne un point de vue plus épistémologique sur la nature des résultats obtenus par une investigation graphique des données. On a vu à plusieurs reprises évoqué le fait que l'observation visuelle des données n'est pas régie par les mêmes mécanismes que le calcul, et que là réside justement sa force face à des phénomènes complexes. Le problème se pose alors à l'inverse face aux difficultés d'évaluation des méthodes d'investigation, et au passage à une autre forme de rationalisation des phénomènes. Autrement dit, « dessiner n'est pas démontrer », du moins pas autant que le permettent les mesures quantitatives, les tests statistiques et les comparaisons entre différentes méthodes.

Les arguments pour la visualisation sont, le reconnaissent Fekete *et al.* (2008), peu adaptés aux modèles dominants de l'évaluation scientifique. Ces arguments se basent soit sur des expérimentations cognitives et perceptuelles, font appel à des situations précises dans lesquelles seule la visualisation a permis de mettre au jour des connaissances, ou encore sur le succès de certaines méthodes que les utilisateurs plébiscitent.

Au-delà des travaux plus techniques de la visualisation des données, le champ récent des *Visual Analytics* (Keim *et al.*, 2008) étudie plus largement les questions méthodologiques et épistémologiques de l'investigation de données massives et complexes, notamment en étudiant l'articulation de ces méthodes visuelles avec les méthodes plus classiques (analyses statistiques et fouille de données, nous y reviendrons dans les prochains chapitres). Le point de vue d'une utilisation plurielle de méthodes d'investigation est au centre des réflexions sur le rôle de la visualisation, comme le résument ici Fekete *et al.* (2008) :

> *Is there a competition between confirmatory, automated and exploratory methods ? No, they answer different questions. When a model is known in advance or expected, using statistics is the right method. When a dataset becomes too large to be visualized directly, automating some analysis is required. When exploring a dataset in search of insights, information visualization should be used, possibly in conjunction with data mining techniques if the dataset is too large.*
>
> (Fekete *et al.*, 2008)

5.3.5 Implications pour la linguistique et le TAL

Je terminerai ce chapitre par quelques conclusions sur les implications de ces avancées récentes pour la linguistique empirique et le TAL.

Il apparaît important de se doter d'outils qui soient suffisamment génériques pour aborder les différents besoins et qui correspondent à la grande variété des données langagières. Comme on l'a dit, ces outils doivent permettre la multiplication des points de vue et permettre l'interaction avec les utilisateurs.

Le savoir-faire en visualisation des données n'est pas de ceux qui se transmettent et se formalisent facilement. On pourrait faire le parallèle avec la grande technicité des spécialistes de l'imagerie médicale, qui doivent atteindre, en plus d'une maîtrise technique des différentes méthodes (et des appareils qui les opérationnalisent), une expertise qui va leur permettre de guider leur recherche d'une pathologie particulière, bien souvent en multipliant les sources d'information et donc les méthodes de visualisation. Ce type d'expertise ne peut sans doute pas se faire sans atteindre une masse critique de situations connues et partagées dans la communauté, seule façon de capitaliser les connaissances pratiques en l'absence d'une méthode universelle de référence.

Comme le souligne Hurter (2010), le domaine de la visualisation de l'information est organisé en communautés, qui se regroupent et s'identifient autour d'objets d'études communs, et qui possèdent bien souvent des espaces référentiels identifiés (cartographie, ingénierie mécanique, médecine, transports, sociologie des réseaux, etc.). Ces communautés permettent la factorisation des coûteux efforts de développement d'outils et l'appropriation de nouvelles méthodes adaptées à leurs données et à leurs besoins. C'est déjà le cas en TAL dans certains domaines plus spécialisés, comme celui de la visualisation des collections de documents, notamment dans le cadre de la recherche d'information pour la présentation des résultats, et pour l'exploration de corpus. Des techniques très sophistiquées de visualisation sont développées, qui font généralement appel à trois dimensions. La disposition des documents dans l'espace se base généralement sur des considérations de proximité thématique, calculée à partir du partage de formes lexicales, sur des méta-données ou des relations explicites comme les hyperliens (voir le panorama tracé par Fang *et al.* (2009)). D'autres travaux s'intéressent plus spécifiquement aux grands documents, en se basant sur leur structure (Jacquemin *et al.*, 2005).

Un récent signal positif semble indiquer que la linguistique en tant que discipline est prête pour la prochaine étape, puisque va se tenir en 2012 une conférence spécifiquement dédiée à ces questions (*AVML : Advances in Visual Methods for Linguistics*[6]).

En attendant cette maturation, il est clair que la seule visualisation ne peut satisfaire aux exigences établies par la communauté de la linguistique sur données. Il reste en effet difficile de se contenter, comme je l'ai fait ci-dessus, de montrer par un dessin, aussi précis et rigoureux soit-il dans sa construction, un résultat qui ait une valeur scientifique suffisante. À l'heure actuelle, d'autres modes de preuve sont nécessaires, au premier rang desquels se trouvent les méthodes statistiques. Les exigences sont d'autant plus radicales en TAL, puisque la validation d'une méthode de traitement des données doit nécessairement se faire par une comparaison rigoureuse avec d'autres méthodes, sur des données calibrées et mesurée par des instruments reconnus.

C'est dans cet esprit que je vais donc, au prochain chapitre, aborder la question de l'utilisation des méthodes statistiques.

6. Voir le site Web de la conférence : `http://avml2012.wordpress.com/`

Chapitre 6

Analyse statistique des données langagières

Avant tout je tiens à préciser que je ne me considère absolument pas comme un spécialiste des questions abordées dans ce chapitre : bien que ma formation de base soit scientifique, avec un certain bagage en mathématiques, elle ne comprenait que très peu de notions dans ce domaine que j'ai donc dû aborder essentiellement en autodidacte. Que les statisticiens me pardonnent donc pour les biais et raccourcis (pour ne pas dire les erreurs) dont font preuve à la fois mon travail et la présentation que j'en fais ici.

La montée en puissance des méthodes statistiques en TAL et en linguistique est désormais une chose acquise, et elles font désormais partie des exigences disciplinaires. Que les techniques de TAL aient évolué globalement dans le sens de l'élaboration de méthodes qui cherchent à exploiter les grandes masses de données maintenant disponibles n'est qu'un des aspects de cette statisticisation (on le verra plus en détails dans la dernière partie de ce mémoire). L'exploitation de données dans les travaux de linguistique descriptive est, elle aussi, de plus en plus une utilisatrice de concepts et de mesures issus de la statistique. Que ce soit à travers la lecture des publications en linguistique de corpus, ou des demandes explicites formulées dans des relectures d'articles soumis, on sent bien que le niveau minimal exigé dans la discipline a nettement monté. Là où il y a quelques années on pouvait se contenter d'une quantification simple (quelques fréquences dans un tableau par exemple), des tests de significativité sont de plus en plus demandés. Pour des approches plus classiques en TAL (qui n'entrent pas dans la catégorie des systèmes par apprentissage notamment, ces dernières étant logiquement concernées au premier chef), l'évaluation et la comparaison quantifiée et validée statistiquement est également devenue obligatoire à travers la pression sociale des journaux et conférences.

Le côté positif est un intérêt et une bien meilleure connaissance des techniques de base par un nombre croissant de chercheurs, ce qui peut, bien entendu, éviter de nombreuses déconvenues et des utilisations peu heureuses de certaines notions, menant dans le pire des cas à des résultats erronés. Le côté négatif est bien entendu l'énergie nécessaire pour s'approprier et mettre en place ces calculs parfois jugés superflus et le manque cruel de documents didactiques abordables pour les bonnes volontés. Bien souvent, les formations en statistique proposées aux linguistes viennent des sciences humaines (essentiellement la sociologie, la psychologie, la géographie et l'histoire) auxquelles elles se sont depuis longtemps intégrées et même adaptées et répondent aux exigences de ces disciplines. Toutefois, quelques ouvrages récents permettent d'aborder plus efficacement ce tournant : je citerai Baayen (2008) (quoiqu'un peu trop ardu et

avec une pente initiale très raide, plus adapté à la psycholinguistique dont l'auteur est issu) et Gries (2009) (beaucoup plus abordable). Les deux ont en commun de se baser sur l'excellent logiciel R [1] qui a le double avantage de couvrir l'ensemble des besoins et d'être libre.

Quoiqu'il en soit, j'ai donc il y a quelques années commencé à utiliser des méthodes statistiques, pour mes propres travaux mais aussi en tant que « consultant » local pour l'ERSS. Fort heureusement, ce deuxième aspect est désormais derrière moi depuis le recrutement récent de collègues bien plus spécialistes que moi (Hélène Giraudo et Basilio Calderone). Je vais donc tenter de résumer ici les différents besoins que j'ai pu identifier, les techniques employées et les résultats obtenus à travers quelques exemples. J'identifierai ensuite une série de questions méthodologiques sur l'utilisation de ces méthodes, ainsi que des pistes à explorer. Ce chapitre n'a pas vocation à être un mini-manuel de statistique pour la linguistique, mais à montrer comment ces méthodes s'inscrivent dans un travail d'investigation méthodique des données langagières. Je m'abstiens notamment d'y exposer des formules et de détailler les mécanismes des méthodes calculatoires, en insistant à l'inverse sur les informations que ces techniques apportent, et sur leur articulation avec les questionnements scientifiques. Cependant, je vais tout de même utiliser un mode nettement plus didactique que dans les autres chapitres de ce mémoire ; je pense en effet que nombre de collègues et d'étudiants ont plus de difficultés par rapport aux questions et aux techniques statistiques que pour les autres aspects liés à l'analyse et à l'interrogation des données langagières (quand bien même ces autres aspects, comme les annotations de corpus, sont techniquement plus complexes).

6.1 Quelles questions sur quelles données ?

Je vais commencer ici par exposer les différents objectifs visés par une analyse statistique, puisque ce seront ces besoins qui organisent la présentation des techniques et des exemples qui suivront dans ce chapitre. J'aborderai également la question de la présentation des données elles-mêmes, et quelques points de vocabulaire.

Les principaux exemples sur lesquels je vais m'appuyer pour le début de ce chapitre concernent l'exploration et l'analyse quantitative des données issues de la campagne d'annotation des structures énumératives dans le projet Annodis (voir (Péry-Woodley *et al.*, 2009) et la présentation des structures en section 5.1.2, page 114), mais je présente par la suite des travaux issus d'autres contextes.

6.1.1 Types de questions

La prédominance déjà évoquée des méthodes statistiques s'explique entre autres par la grande diversité des questionnements pour lesquels elles apportent des solutions et des pistes de réponses. On peut regrouper ces différentes questions en trois grandes catégories, dont la diversité interne est elle aussi très importante.

6.1.1.1 Dégager les grandes tendances et les cas à part

L'objectif le plus simple est celui qui concerne une présentation générale d'un jeu de données langagières. Qu'il s'agisse d'un corpus, ou d'une collection d'unités recueillies, tout travail d'exploitation doit nécessairement débuter par un examen et une présentation générale

1. `http://www.r-project.org/`

des caractéristiques de ces données. Quand il s'agit de caractéristiques externes (sources, type, etc.), il est en effet primordial de donner la distribution de celles-ci sur l'ensemble des données assemblées, pour expliciter les choix de sélection ou constater la présence et la répartition des différentes catégories afférentes. Dans le cas très fréquent des données annotées, les caractéristiques ajoutées doivent elles aussi être présentées de façon synthétique.

Ces observations globales (sur l'ensemble du jeu de données étudié) des différentes caractéristiques permettent dans un premier temps d'avoir un aperçu des grandes tendances observables. Il est ainsi possible, en utilisant des mesures ou des représentations simples, d'exprimer l'étendue d'un phénomène, sa diversité, ou encore d'avoir le profil moyen (ou typique) des unités qui composent la collection.

Pour les données du projet Annodis, ceci concerne par exemple la présentation du nombre, de la taille, et de la répartition dans les différents types de texte des structures énumératives qui ont été annotées.

Un autre point important et faisant appel aux mêmes techniques concerne le repérage des individus particuliers, qui peuvent être distingués très facilement si une (ou plusieurs) de leurs caractéristiques les fait s'éloigner notablement du profil moyen ou des grandes tendances générales. On verra dans les exemples que ces situations peuvent correspondre au repérage d'erreurs dans le processus de sélection ou d'annotation des données (et permettre d'améliorer les données assemblées), ou bien permettre d'identifier des cas particulièrement intéressants au sein de la masse.

On verra sur les données du corpus Annodis que cette opération permet de repérer des erreurs d'annotation (des structures incomplètes par exemple) ou des cas-limites (des structures ayant un très grand nombre d'items).

6.1.1.2 Comparer des données et étudier la variation

Les descriptions synthétiques peuvent s'appliquer à des sous-ensembles de la collection, comme c'est très souvent le cas en linguistique de corpus. Il est en effet très courant de vouloir observer (et mesurer) des variations entre deux sous-ensembles de données, par exemple en étudiant un même phénomène linguistique dans différents environnements (comme les types de textes au sens large). La première approche consiste alors à comparer les caractéristiques des données dans différentes configurations : les outils statistiques donnent ainsi des solutions directes pour mesurer l'existence et l'ampleur de différences dans ces tendances et/ou dans les distributions des phénomènes étudiés.

Sur les données du projet Annodis, on verra par exemple si (et comment) les structures énumératives varient (en taille, nombre, etc.) entre les différents types de textes du corpus.

Cette étude de la variation est également très courante pour évaluer un traitement automatique des données, qu'il s'agisse d'une annotation automatique ou du calcul d'une caractéristique des données. Il est important dans ce cas de comparer les résultats obtenus à une situation de référence (par exemple une annotation manuelle) ou à une autre méthode de calcul. Dans les deux cas, il s'agit d'avoir un point de vue quantifié sur l'efficacité de la méthode de calcul, et c'est bien entendu un enjeu majeur pour tous les travaux en TAL.

6.1.1.3 Découvrir des relations entre les caractéristiques

Découvrir que les caractéristiques d'un objet linguistique varient d'une configuration à une autre est en fait un cas particulier d'un ensemble de phénomènes que les mesures statistiques

permettent de mettre au jour. Dans le cas plus général, elles consistent à mesurer la liaison entre différentes caractéristiques, et donner ainsi des indices pour une meilleure compréhension des mécanismes langagiers sous-jacents.

Cette notion de liaison est fondamentale pour une grande partie des outils statistiques : elle consiste en une tendance observable à travers une collection de données qui associe une caractéristique à une autre (ou à plusieurs autres) sur la base de variations d'un individu à un autre. Elle se traduit intuitivement par l'influence que l'on attribue à une caractéristique sur une autre, et est fondamentale dans la compréhension de phénomènes complexes. Il s'agit généralement de l'objectif principal de tout travail faisant appel à des données.

Sur les données du projet Annodis, on mesurera par exemple en quoi la taille d'une structure a une influence sur ses autres caractéristiques (structure, nombre d'indices, type, etc.)

Dans certains travaux, ces phénomènes peuvent être bien entendu présupposés, provenant d'observations isolées, d'intuitions ou de modèles théoriques. Les méthodes statistiques permettent alors de tester une telle hypothèse en quantifiant les phénomènes. Dans d'autres situations, il peut s'agir d'une approche plus exploratoire, dans laquelle l'existence de telles liaisons est le résultat d'un calcul systématique. Dans les deux cas, les méthodes statistiques fondamentales sont les mêmes.

6.1.1.4 Généraliser des hypothèses

Le dernier objectif est celui qui explique l'importance des statistiques en tant qu'argument dans l'établissement d'un fait nouveau observé sur des données. Il s'agit de la capacité de partir d'un phénomène observé sur une quantité réduite de cas, et d'envisager son extension à un cas plus général. Un pan majeur du développement des méthodes statistiques concerne la mesure de la plausibilité de ce type de généralisation. Les mécanismes mis en œuvre pour ce faire sont assez complexes, et concernent notamment la notion d'échantillonnage des données. Un échantillon est un ensemble fini de données individuelles que l'on considère comme extrait d'un ensemble bien plus grand des données existantes du même type. C'est très souvent le cas lorsque l'on se donne un corpus d'étude en sélectionnant une série de textes d'un type particulier (ou même, quand on prend des textes tout venants sans poser de contraintes précises). Toute observation d'un phénomène dans ce corpus peut être due à l'existence d'un principe général (qui se retrouve dans tout texte du type étudié) ou bien à une configuration très spécifique aux données sélectionnées. La statistique inférentielle a explicitement pour but de faire la part des choses, et de proposer une estimation (sous forme de probabilités) du danger qu'il y aurait à généraliser les résultats à des données extérieures au corpus d'étude.

Dans le cadre du projet Annodis, il s'agira de vérifier si l'on peut, au regard des seules données annotées dans le corpus, tirer des conclusions valables pour toute structure énumérative.

Il existe ainsi une panoplie de mesures classiquement utilisées pour mesurer cette plausibilité de la généralisation (également appelée significativité). Du point de vue pratique, il s'agit de calculs spécifiques permettant d'obtenir au final une réponse binaire à la question de la validité d'une hypothèse à partir des données. C'est généralement ce genre de mesure qui est attendue dans toute publication prenant appui sur une observation de données. Je reviendrai plus longuement sur les problèmes posés par ce type de pratique dans la discussion finale de ce chapitre.

6.1.2 Données de départ

L'utilisation de méthodes statistiques nécessite de disposer de données organisées de la façon suivante : une collection d'individus qui correspondent aux unités à étudier, chacun de ces individus étant décrit par un ensemble de caractéristiques (variables). Le tout prend donc la forme générale d'un tableau (comme 6.1) dans lequel les individus sont représentés par les lignes et les variables par les colonnes. La taille de ce tableau va dépendre du nombre d'objets étudiés mais aussi de la variété des informations dont on dispose sur ceux-ci.

Cette situation est celle que l'on obtient naturellement lorsque l'on effectue une tâche d'annotation, que les individus soient sélectionnés dès le départ ou obtenus par un processus d'identification en corpus. Les caractéristiques utilisées pour qualifier ces individus vont dépendre de leur nature, de l'étude spécifique, et des différents phénomènes liés que l'on va prendre en compte.

A l'issue de la phase d'annotation manuelle des structures discursives du projet Annodis nous disposons de 849 structures énumératives identifiées dans le corpus par les annotateurs (voir section 5.1.2, page 114 pour plus d'informations sur l'annotation et sur les structures elles-mêmes). La table 6.1 montre un extrait de la matrice obtenue lorsque l'on a calculé et rassemblé les principales caractéristiques de ces structures.

Id	Corpus	Texte	Annot.	Mots	Par.	Items	Amorce	Clôture	Indices	Type
S_1890	CMLF	CMLF_dal	A	344	1	3	Oui	Non	4	T2
S_0656	CMLF	CMLF_dal	A	623	1	5	Oui	Non	8	T2
S_4703	CMLF	CMLF_dal	A	71	0	2	Oui	Non	4	T4
S_8292	GEOPO	geopo_30	C	235	0	7	Oui	Non	12	T4
S_4637	GEOPO	geopo_30	C	77	1	5	Oui	Non	12	T2
S_5221	GEOPO	geopo_30	C	67	1	4	Oui	Non	7	T2
S_9546	CMLF	CMLF_dal	A	2085	25	3	Oui	Non	7	T1
S_6140	CMLF	CMLF_colas	B	146	0	2	Oui	Oui	5	T4
S_5359	CMLF	CMLF_dal	A	162	1	2	Non	Non	2	T3

TABLE 6.1 – Extrait de la table de données sur les structures énumératives du projet Annodis

Les caractéristiques qui décrivent ces objets viennent de différentes sources :
- des caractéristiques liées à l'origine de ces données : un identifiant unique pour chaque SE (*Id*), le texte où elles ont été trouvées (*Texte*), le sous-corpus auquel ce texte appartient (*Corpus*[2]) et enfin l'identifiant de l'annotateur (*Annot.*) ;
- des caractéristiques résultant de l'annotation et décrivant la structure elle-même : sa taille en nombre de mots (*Mots*) et de paragraphes (*Par.*), sa constitution : présence ou non d'une amorce (*Amorce*), d'une clôture (*Clôture*), nombre d'items (*Items*), le nombre et la nature des indices qui lui ont été associés (*Indices*) ;
- des résultats d'une interprétation déjà réalisée à ce stade, comme dans le cas d'Annodis le type de structure (parmi les quatre identifiés : *T1* = SE composée de sections titrées, *T2* = liste à puces, *T3* = multi-paragraphique, *T4* = intra-paragraphique.

2. Trois sous-corpus ont été utilisés : des articles de linguistique du premier congrès mondial de linguistique française (CMLF), des articles de géopolitique de l'Institut Français de Relations Internationales (GEOPO) et des articles de la Wikipedia (WIKI).

Sur le plan technique, ces caractéristiques varient également en fonction de la nature des valeurs : on distinguera ici simplement les variables numériques (taille, longueur, nombre d'items, nombre d'indices, etc.) des variables nominales (type, sous-corpus d'origine, identité de l'annotateur) ou booléennes (présence/absence d'une amorce, d'une clôture). On verra que la prise en compte de cette nature des variables est essentielle car elle implique l'utilisation de techniques totalement différentes pour leur description et les mesures qui les impliquent.

6.2 Description d'un jeu de données

Cette section concerne donc les moyens de répondre aux premières questions évoquées ci-dessus : quelles sont les grandes tendances observables dans le jeu de données étudié, y a-t-il des individus particuliers qui se dégagent de ces grandes tendances, et quelles sont les principales variations observables au sein de la collection ?

Ces différentes approches nécessitent de mettre en œuvre les méthodes les plus simples de la statistique, à savoir les mesures de statistique descriptive. Les premières que je vais présenter concernent l'étude d'une variable unique (monovariée), et les secondes vont étudier les interactions entre deux variables (bivariées).

6.2.1 Tendances générales et distributions

Pour les méthodes monovariées, tout comme pour les autres techniques statistiques, il existe de grandes différences en fonction de la nature de la variable étudiée. Je vais donc séparer la présentation en deux sous-parties.

6.2.1.1 Description d'une variable nominale : répartition des structures énumératives (SE) par type

Les méthodes statistiques donnent peu de possibilités pour décrire une variable non-numérique (qu'elle soit nominale ou booléenne). La seule façon de synthétiser une telle variable (par exemple le type de SE) est d'établir les effectifs (nombre d'individus) pour chaque valeur possible. On aura donc simplement comme synthèse le nombre de SE de chaque type, exprimable également en pourcentages. Ces informations ne permettent que d'observer d'éventuels déséquilibres entre les classes définies par une telle variable, ou d'identifier la classe la plus fréquente. Par exemple, le calcul présenté dans la table 6.2 permet de vérifier que les quatre types de SE définis par la typologie proposée sont équilibrés, mais que le type 4 (intra-paragraphique) est le plus fréquent, et le type 1 (sections titrées) le plus rare.

Type	T1	T2	T3	T4
Effectifs	104	223	194	308
%	12,25	26,27	22,85	38,63

Table 6.2 – Répartition des différents types de structures énumératives

Ce calcul s'applique également aux variables booléennes et permet ainsi de connaître le taux de SE qui ont une amorce et un clôture. Plusieurs représentations graphiques sont utilisables, notamment les diagrammes à secteurs (camemberts) ou à barres.

6.2.1.2 Description d'une variable numérique : nombre d'items des SE

Il existe, au contraire des variables nominales, de nombreuses mesures permettant de synthétiser une variable dont les valeurs sont des nombres. Ces mesures ont pour but d'identifier la tendance (valeur centrale) et la distribution (dispersion) de la liste des valeurs correspondant à une caractéristique à l'échelle de l'ensemble des données.

Les mesures de valeur centrale sont par exemple la moyenne et la médiane. Les mesures de dispersion classiques sont les valeurs extrêmes (minimum et maximum), la variance et l'écart-type.

Par exemple, pour ce qui concerne le nombre d'items des structures énumératives, on observe les valeurs présentées dans la table 6.3.

Minimum	Maximum	Moyenne	Médiane	Variance	Ecart-Type
1	48	3,38	3	6,78	2,60

TABLE 6.3 – Caractéristiques principales du nombre d'items des structures énumératives

Plusieurs informations importantes émergent de ces calculs. Premièrement, le minimum de 1 indique clairement qu'une SE au moins n'a pas été correctement annotée : la définition de la tâche d'annotation indique qu'une SE doit contenir au moins deux items. Ensuite, le maximum de 48 indique qu'une SE au moins correspond à un phénomène inattendu, ce nombre d'items est très élevé, et peut correspondre à un cas limite du phénomène. La moyenne et la médiane sont proches, mais la moyenne est tout de même supérieure, ce qui indique que plusieurs SE ont des valeurs élevées qui étirent cette moyenne vers le haut. L'écart-type (qui n'est que la racine carrée de la variance) indique en effet qu'il existe d'importantes variations.

Plusieurs conséquences vont découler de ces constatations : premièrement, plusieurs SE vont être simplement supprimées de la collection car elles ne correspondent pas à la définition (on supprimera toutes celles qui ont moins de 2 items). Ensuite, les SE qui ont un nombre d'items très élevé vont devoir être examinées de plus près, individuellement.

Pour avoir un aperçu plus global de la distribution de cette variable, le plus simple est de faire appel à une représentation graphique, comme un histogramme qui associe à chaque valeur (ou intervalle de valeurs) d'une variable le nombre d'individus correspondants.

L'histogramme de la figure 6.1 permet de déduire plusieurs points :

- il existe effectivement plusieurs SE dont le nombre d'items est très important, ce qui les isole du reste de la collection. Les valeurs très élevées (une seule SE pour chacune) sont : 16, 19, 20, 23 et 48 ;
- les nombres d'items les plus représentés sont 2 (370 SE), puis 3 (225 SE) ;
- la courbe dans son ensemble a une forme qui indique une décroissance très rapide avec un léger étalement à droite.

Le premier point permet d'identifier les individus particuliers, qui s'éloignent notablement du reste de la collection. On appelle classiquement ces individus des *outliers*, et il est important de les identifier le plus tôt possible. Dans la plupart des cas, il est même nécessaire de les écarter de la collection pour la suite de l'analyse, puisque leurs valeurs ont tendance à fausser les résultats. Concrètement, ces SE correspondent à des suites chronologiques d'événements parfois très longues : seul un des trois annotateurs a jugé bon de les identifier, et cette observation permet notamment d'affiner la description de la structure visée, puisqu'elle met le doigt sur une ambiguïté dans la frontière entre cette structure et d'autres objets textuels.

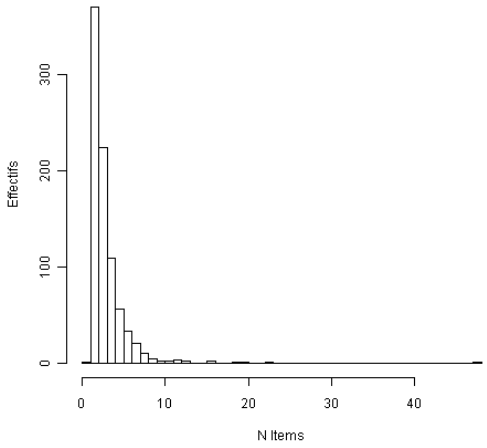

FIGURE 6.1 – Histogramme du nombre d'items par structure énumérative

Le second point est l'information synthétique la plus pertinente à ce stade : environ la moitié des structures énumératives comprennent le nombre minimal d'items.

Le troisième et dernier point est très important pour la suite : il s'agit de l'identification de la correspondance (ou non) de la distribution de la variable observée avec un des grands types de distributions sur lesquelles se basent les modèles de la statistique. Ces distributions sont des idéaux mathématiques (appelées *lois*) issus de la théorie des probabilités, et qui décrivent le comportement de variables numériques par des formules précises. De nombreuses méthodes statistiques sont basées sur ces lois, en ce sens que le comportement d'une variable observée empiriquement pourra être comparé à celles-ci, et que certaines des mesures que l'on utilise pour interpréter des données ont été calibrées spécifiquement en se basant sur ces modèles mathématiques. Par exemple, une grande partie des méthodes plus complexes d'analyse statistique des données sont basées sur la loi normale (une distribution dont la représentation graphique a la forme d'une cloche, symétrique, autour de la valeur moyenne). Ainsi, savoir si une variable numérique spécifique semble (ou non) correspondre à cette loi aura des conséquences dans le choix de certaines techniques, notamment quand il s'agira de mesurer et de quantifier ses interactions avec d'autres variables.

On voit notamment dans la figure 6.1 que le nombre d'items ne présente absolument pas cette forme caractéristique de cloche. Dans certains cas toutefois, il est aisément possible de transformer les valeurs pour obtenir une autre forme de distribution. Un des outils les plus simples pour ce faire est le logarithme, comme le montre en figure 6.2 l'histogramme de distribution du nombre de mots d'une SE, avant et après application d'un logarithme.

La transformation logarithmique permet de séparer les valeurs resserrées dans la partie basse du spectre, et au contraire de rapprocher les valeurs étalées vers la droite. Nous verrons plus loin que cette transformation est donc toujours souhaitable pour l'observation de cette variable. Ce type d'opération simple sur les données pour faciliter leur traitement fait partie d'un ensemble de savoir-faire propres aux analyses statistiques, qui dépendent fortement du type de données étudiées.

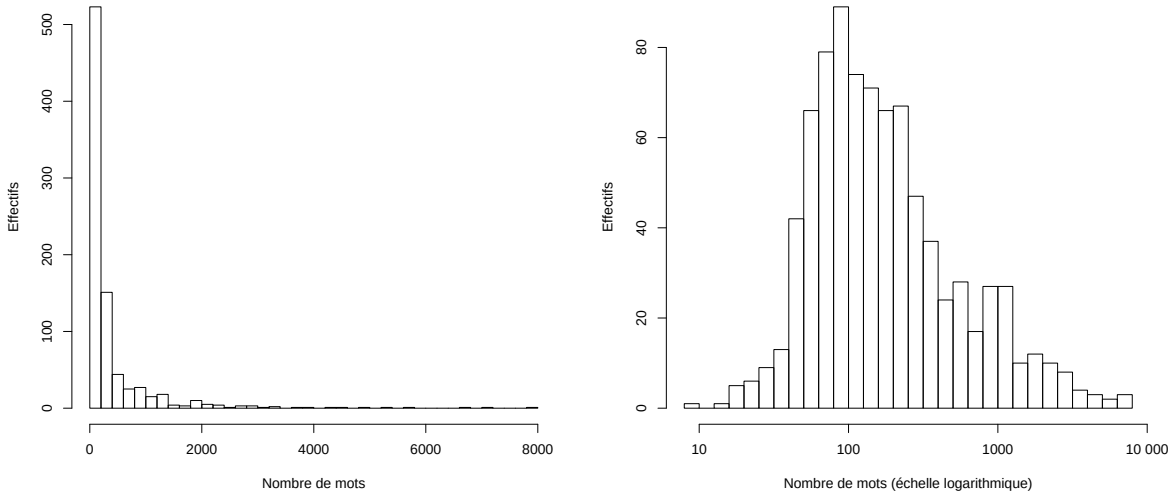

FIGURE 6.2 – Distribution du nombre de mots des SE, avec et sans transformation logarithmique

L'étude des variables prises isolément ne peut pas aller beaucoup plus loin que ce qui a été montré précédemment. Les questions les plus intéressantes vont correspondre à l'étude des liaisons entre deux variables.

6.2.2 Comparaison de caractéristiques

Les questions plus complexes que l'on va aborder concernent des comparaisons au sein de la collection de données.

Un ensemble de questions émergent naturellement : les SE sont-elles plus nombreuses, ou ont-elles des caractéristiques différentes dans un sous-corpus par rapport à un autre ? En quoi les différents types de SE se distinguent-ils en termes de structure ou de taille ? N'y a-t-il pas eu un biais dans les données par l'appel à plusieurs annotateurs ?

Dans tous ces cas, la méthode consiste à *croiser* deux variables et à observer la répartition des valeurs correspondantes. En fonction de la nature des variables, les méthodes de représentation vont toutefois être totalement différentes.

6.2.2.1 Croisement de variables nominales : répartition des SE par corpus et par type

Le premier cas à envisager est celui où l'on croise deux variables nominales, donc un découpage des individus suivant deux caractéristiques distinctes. Comme c'est le cas lorsque l'on étudie une seule variable nominale, le calcul des effectifs est un passage obligé. On utilise pour ce type d'étude un tableau croisé, ou table de contingence comme celle de la table 6.4.

En plus des effectifs bruts, les pourcentages de répartition sont souvent utiles pour comparer les profils. Ici, ce sont les profils de ligne qui ont été calculés, pour permettre d'identifier ce qui différencie les corpus en termes de répartition de SE. Des différences apparaissent clairement, par exemple le fait que les structures de type 1 (sections titrées) sont bien plus

Corpus/Type	T1	T2	T3	T4	Total
CMLF	24 (9%)	61 (23%)	70 (27%)	108 (41%)	263
GEOPO	16 (6%)	32 (13%)	55 (22%)	151 (59%)	254
WIKI	64 (20%)	124 (38%)	67 (21%)	68 (21%)	323
Total	104 (12%)	217 (26%)	192 (23%)	327 (39%)	840

TABLE 6.4 – Table de contingence de la répartition des types de SE par sous-corpus

fréquentes dans la Wikipedia que dans les autres corpus. Toutefois, ces indications restent difficiles à interpréter directement, notamment lorsque le nombre de modalités des variables implique des tableaux de contingence plus volumineux (i.e. quand le nombre de modalités des variables est important).

6.2.2.2 Croisement d'une variable nominale et d'une variable numérique : différences de taille entre les types de SE

Lorsque deux variables de types différents sont examinées conjointement, cela revient à comparer plusieurs séries de valeurs numériques. La variable nominale est vue comme un découpage de la collection en plusieurs sous-classes, chacune de ces sous-classes permettant donc d'identifier une série de valeurs numériques. Par exemple, si l'on croise le type de structures énumératives avec la taille de celles-ci en nombre de mots, on obtient 4 séries de valeurs numériques à comparer. On peut bien entendu utiliser les mêmes méthodes de description que vues précédemment pour chacune de ces 4 séries (minimum, maximum, moyenne, etc.) et les résumer comme dans la table 6.5.

Nb de mots/Type	Minimum	Maximum	Moyenne	Médiane	Ecart-Type
T1	231	7831	1671,0	1206	1362,3
T2	8	1541	166,7	115	179,6
T3	290	7174	401.2	258	592,8
T4	13	730	108.8	131	68,8

TABLE 6.5 – Comparaison du nombre de mots par type de structure énumérative

On peut y voir que les nombres moyens de mots sont effectivement très différents ; cependant les écarts-types indiquent également une grande dispersion. Pour observer plus globalement la distribution des tailles de SE suivant leur type, le plus efficace est à ce stade de recourir à une représentation graphique. Les *boxplots* ou *boîtes à moustaches* proposent une représentation synthétique d'un ensemble de valeurs numériques, en indiquant sur un même schéma les valeurs extrêmes, la médiane (trait horizontal gras) et les deux limites autour de la médiane qui contiennent la moitié des valeurs de la série (boîte). La figure 6.3 présente ainsi les 4 boxplots correspondant au nombre de mots pour chaque type de SE. Toutefois, afin de faciliter la représentation, c'est le logarithme du nombre de mots qui a été utilisé.

L'absence de recouvrement des boîtes, notamment en ce qui concerne les types 1 et 3 indique bien que la taille des SE varie de façon importante avec leur type, et que les SE de

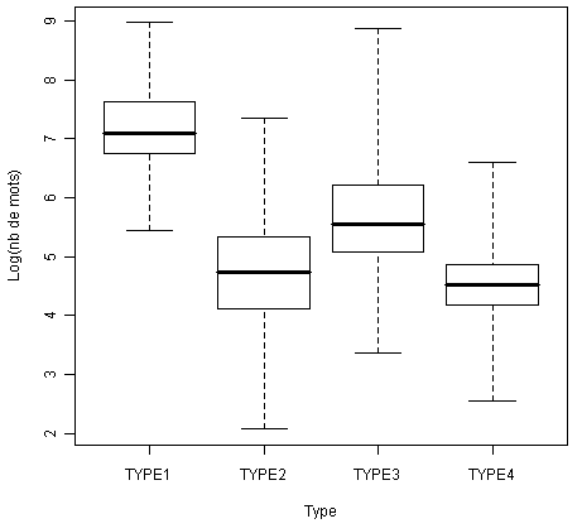

FIGURE 6.3 – Boîtes à moustaches : nombre de mots (log) par type de SE

type 1 et 3 sont nettement plus longues. Les types 2 et 4 semblent avoir des tailles similaires, mais celles du type 2 montrent une variation plus importante.

6.2.2.3 Croisement de deux variables numériques : rapport entre les nombres d'indices et d'items dans les SE

Croiser deux variables numériques revient à isoler pour chaque individu deux valeurs correspondantes : on obtient alors une liste de couples qu'il va falloir comparer à l'échelle de la collection.

Le plus simple est, comme d'habitude, de représenter ces données par un graphique afin d'observer globalement le comportement. Par exemple, si l'on souhaite comparer le nombre d'indices de chaque SE avec son nombre d'items, on va positionner sur un graphique (appelé diagramme de dispersion, ou *scatterplot*) des points représentant chacun une SE : l'abscisse de ces points sera le nombre d'items et l'ordonnée le nombre d'indices. L'ensemble de la collection forme donc un nuage de points comme celui de la figure 6.4. Pour mieux représenter ces données, une perturbation aléatoire mineure (*jitter*) a été apportée aux coordonnées, ce qui évite que les points soient trop superposés.

On peut y voir que globalement, le nombre d'indices associés à une SE est d'autant plus important que celle-ci a beaucoup d'items, bien qu'une grosse partie de la collection corresponde à des SE à 2 ou 3 items.

6.3 Mesurer des phénomènes et prouver leur existence

On a pu remarquer une grande similitude entre les représentations graphiques (et les mesures numériques simples) des sections précédentes et les approches visuelles présentées au chapitre précédent. Toutes deux permettent de forger un point de vue global sur certaines

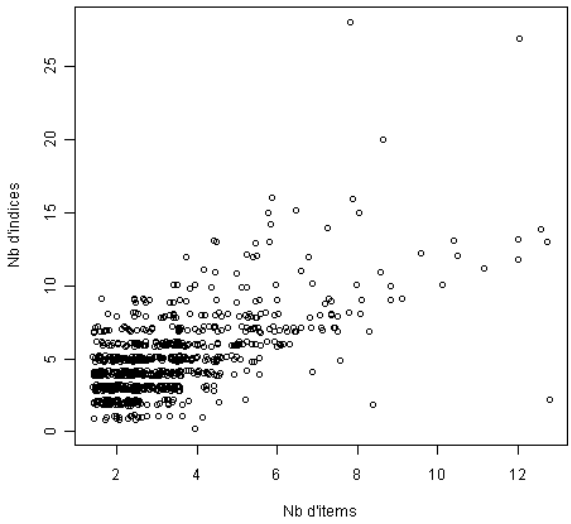

FIGURE 6.4 – Diagramme de dispersion du nombre d'indices et du nombre d'items des SE

caractéristiques des données, et de guider une analyse plus détaillée. Toutefois, comme on l'a dit, ces modes d'approche se heurtent rapidement à un besoin d'objectivation de ces observations, et donc d'une forme de quantification des phénomènes mis au jour. C'est cet objectif de mesure des associations entre variables et des différences entre des sous-ensembles des données qui vont être présentés ici.

6.3.1 Mesurer la dépendance entre deux caractéristiques

La statistique propose ainsi un nombre (important) de mesures de l'intensité de la dépendance (ou liaison) observée entre deux variables. Cette notion de dépendance correspond simplement aux constatations que l'on a pu faire à l'issue de l'observation des exemples précédents : il semblerait que les sous-corpus contiennent des SE de types différents, que ces mêmes types de SE correspondent à des tailles différentes, et que le nombre d'indices soit proportionnel au nombre d'items.

6.3.1.1 Mesure du Khi-deux entre deux variables nominales

A partir d'un tableau de contingence comme celui de la table 6.4, la mesure de la liaison entre les deux variables impliquées nécessite d'exprimer plus précisément le phénomène recherché. S'il y a une liaison entre le type de SE et le sous-corpus où celles-ci ont été identifiées, cela signifie que les répartitions des SE par type vont être différentes entre chaque sous-corpus. Si, par contre, une telle liaison n'existe pas, alors les répartitions devraient être à peu près identiques. Une mesure de cette liaison va donc être exprimé par la différence entre les effectifs observés (ceux de la table 6.4) et ceux qui *auraient été obtenus* si l'on considère qu'il n'y a aucun impact du corpus sur le type de SE. Cette configuration supposée correspond à ce que la statistique appellent *l'hypothèse nulle*, ou l'hypothèse d'indépendance entre les deux variables.

Il est en fait très simple d'obtenir ces valeurs : il suffit de projeter sur chaque combinaison de valeurs les répartitions globales (indiquées par les sommes des lignes et des colonnes) pour obtenir le nombre de SE de chaque type que l'on s'attendrait à observer si la répartition était indépendante du corpus. Cela donne la table 6.6 ; par exemple le nombre *attendu* de SE de type 1 dans le corpus CMLF est obtenu en multipliant 104 (nombre de SE de type 1 dans l'ensemble du corpus) par 263 (nombre de SE dans le corpus CMLF) et en divisant le tout par 840 (nombre total de SE).

Observés	T1	T2	T3	T4	Total
CMLF	24	61	70	108	263
GEOPO	16	32	55	151	254
WIKI	64	124	67	68	323
Total	104	217	192	327	840

Attendus	T1	T2	T3	T4	Total
CMLF	32,6	67,9	60,1	102,4	263
GEOPO	31,4	65,6	58,0	98,9	254
WIKI	40,0	83,4	73,8	125,7	323
Total	104	217	192	327	840

TABLE 6.6 – Effectifs observés et attendus (type de SE versus sous-corpus)

Dès lors, il ne reste plus qu'à comparer ces valeurs deux à deux : on peut ainsi observer s'il y a déficit (la valeur observée est plus petite qu'attendu) ou excédent, ou si les deux valeurs sont proches. Le résultat de cette comparaison peut prendre la forme d'une table similaire à celles de la figure 6.6, ou mieux, celle d'un graphique comme celui de la figure 6.5. Dans cette dernières, la hauteur des barres verticales indique l'ampleur du décalage entre les effectifs observés et attendus, pour chaque association possible entre type de SE et corpus. La largeur des barres est, elle, proportionnelle au nombre (attendu) de chaque cas.

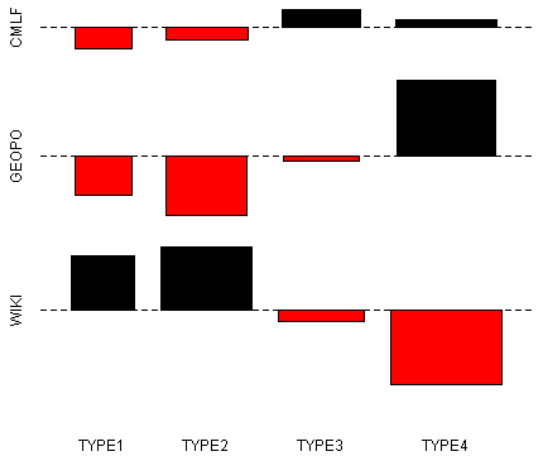

FIGURE 6.5 – Déficits et excédents des types de SE par corpus

On peut voir sur ce graphique qu'il semble y avoir une affinité entre les SE de type 4 et le corpus *Geopo*, à l'inverse des textes de la Wikipedia. Ce dernier corpus privilégie, lui,

les SE des types 1 et 2. Les articles de linguistiques extraits du CMLF montrent, eux, un comportement relativement équilibré par rapport aux deux autres.

En plus de cette représentation graphique, on peut mesurer l'écart global entre les valeurs pour avoir une représentation synthétique de la liaison entre les deux variables croisées. C'est le rôle attribué classiquement à la mesure du χ^2 (khi-deux), que l'on obtient en additionnant pour chaque valeur le carré de l'écart entre observé et attendu divisé par la valeur attendue. Cette valeur globale (ici 118,6), est difficilement interprétable : considérons simplement que plus elle est élevée, plus les écarts sont importants (une valeur nulle indique qu'il n'y a pas de différence entre l'observé et l'attendu). Par contre, la valeur absolue du χ^2 dépend, en plus de l'ampleur des écarts, du nombre d'individus sur lesquels elle a été calculée.

Il existe donc des mesures plus facilement interprétables, comme le coefficient V de Cramer. Cette valeur est un nombre compris entre 0 et 1, 0 étant le cas obtenu lorsqu'il n'y a aucun écart, et 1 lorsque les écarts sont maximaux. Dans le cas des données présentées ici, la valeur pour cette mesure est de 0,27. Si ce coefficient n'est pas plus facilement interprétable que la valeur du χ^2, elle a l'avantage de prendre en considération le nombre d'individus de la collection étudiée, et de normaliser la mesure de la dépendance. Il est donc notamment possible de comparer deux valeurs de ce coefficient, pour comparer deux liaisons (par exemple sur deux collections distinctes, ou deux sous-ensembles de la même collection).

On verra plus loin un autre mode d'interprétation de ces valeurs, dans le cadre des tests statistiques.

6.3.1.2 Coefficient(s) de corrélation entre deux variables numériques

Lorsque la liaison doit être mesurée entre deux valeurs numériques, la mesure la plus simple est celle de la corrélation linéaire. Cette mesure est calculée en prenant en compte les variations conjointes des deux variables à travers la collection d'individus étudiés. Si deux variables sont fortement corrélées positivement, alors un individu qui a une valeur importante pour l'une des deux a de très fortes chances d'avoir également une valeur élevée pour l'autre, et vice-versa pour les valeurs faibles. Dans le diagramme de dispersion de la figure 6.4, on peut voir que les points sont (grossièrement) alignés sur une ligne diagonale : les SE qui ont plus d'items que la moyenne ont également plus d'indices que la moyenne, et les cas contraires sont très rares.

Le coefficient de corrélation de Pearson (noté r) permet de résumer ce type de liaison en une valeur comprise entre -1 et 1. Une valeur négative indique une corrélation inverse (plus un individu est X, moins il est Y) ; la valeur absolue du coefficient est une indication de la force de la corrélation. Une valeur nulle indique l'absence totale de corrélation entre les deux variables. Dans le cas du nombre d'items et du nombre d'indices, la valeur du coefficient de corrélation est de 0,65, ce qui traduit donc bien l'observation faite à la lecture du graphique.

Ce coefficient, très utilisé, est toutefois limité par plusieurs points : premièrement il ne mesure qu'une corrélation *linéaire* (il ne permet d'identifier que les cas où les points du diagramme de dispersion sont alignés sur une droite), et pas des relations plus complexes pouvant exister entre deux variables numériques. Deuxièmement, sa définition et son interprétation (voir plus loin) supposent que les variables étudiées sont distribuées suivant la loi normale, or on a vu que ce n'était pas le cas pour le nombre d'items des SE (ni d'ailleurs pour le nombre d'indices). De ce fait, le coefficient est très sensible aux valeurs extrêmes, ce qui est une raison supplémentaire d'identifier les *outliers* et de ne pas les inclure dans ce genre d'approche, comme cela a été fait ici.

Les alternatives au coefficient de corrélation de Pearson sont notamment le coefficient ρ de Spearman et le τ de Kendall. Tous deux effectuent un calcul similaire, mais en ne prenant en compte que les valeurs relatives entre les individus : au lieu d'utiliser les valeurs absolues ils se basent sur les *rangs* de ces valeurs dans la collection (la valeur la plus élevée aura le rang 1, la suivante dans l'ordre décroissant le rang 2, indépendamment de l'écart entre celles-ci). Ce type de mesure est appelée *non-paramétrique*, en ce sens qu'elle ne suppose pas que les données utilisées aient une distribution spécifique. Par contre, leur interprétation est la même que celle du coefficient de Pearson. Les valeurs suivantes ont été calculées sur nos données : $\rho = 0.57$ et $\tau = 0.47$. Bien que la conclusion soit du même ordre (une corrélation positive moyenne), les scores sont moins élevés car les SE possédant un grand nombre d'items prennent une part moindre dans le calcul.

6.3.1.3 Mesure de la liaison entre une variable numérique et une variable nominale

Les méthodes sont légèrement plus complexes lorsqu'il s'agit de croiser une variable numérique et une variable nominale : le principe est de calculer dans quelle mesure varient les caractéristiques de distribution des séries de valeurs numériques correspondant à chacune des classes induites par la variable nominale.

Une des mesures les plus classiques concerne la comparaison des moyennes des différentes valeurs, c'est ce que fait la valeur t de Student. Une autre approche est de comparer la variance entre les séries de valeurs, c'est ce que font les méthodes qui constituent la famille des mesures d'ANOVA (pour *analysis of variance*). L'idée à chaque fois, pour ces mesures, est de voir si les séries sont similaires ou non, et si les écarts mesurés entre ces caractéristiques sont suffisamment importants. Comme pour la mesure du χ^2, ces calculs comparent les valeurs observées à ce qui serait obtenu si l'on n'avait pas fait de distinction entre les ensembles induits par la prise en compte d'une variable nominale.

Nous n'entrerons pas ici dans le détail des calculs de ces mesures, qui de toute façon sont difficilement interprétables telles quelles. Par contre, elle entrent dans le processus des tests d'hypothèses statistiques que nous allons maintenant voir plus en détail.

6.3.2 Prouver et généraliser : les tests statistiques

Les tests statistiques (ou tests d'hypothèse) sont la finalité des différentes mesures de liaison que nous avons esquissées précédemment. Ces tests permettent une interprétation directe pour pouvoir déduire l'existence ou non d'une liaison (ou dépendance) entre deux variables. Le principe général en est le suivant : on considère tout d'abord que les données étudiées forment un *échantillon* d'un ensemble plus grand de données généralement inaccessible pour l'étude (par exemple l'ensemble des structures énumératives de tous les textes du même type que ceux considérés dans le projet Annodis). Les variations conjointes entre les valeurs de deux variables que l'on observe (différence de moyenne entre deux sous-ensembles, différence de répartition, coefficient de corrélation, etc.) sont donc explicables par deux raisons principales : soit il existe effectivement un principe général qui gouverne le fonctionnement des objets étudiés et qui entraîne les régularités observées, soit tout cela n'est dû qu'aux choix arbitraires (appelons cela naïvement le hasard) qui ont conduit à réunir ces données-là alors que d'autres auraient très bien pu être utilisées dans le même but.

6.3.2.1 Principes

Le rôle d'un test statistique est d'estimer la part de ces deux aspects. Pour ce faire, les méthodes donnent essentiellement une estimation des probabilités de l'erreur qui consisterait à attribuer à une analyse une conclusion de dépendance entre deux variables alors que celle-ci n'est en fait due qu'à un (heureux) choix des données. Cette probabilité est généralement comparée à un seuil que la doxa statistique fixe arbitrairement en fonction des disciplines et des habitudes. Il est convenu pour le type d'études que nous faisons en sciences humaines qu'une probabilité d'erreur inférieure à 5% est admise (les autres seuils classiquement utilisés sont 1% et 0,1%). Ainsi, si une liaison particulière observée sur les données est interprétée par un test statistique comme ayant moins de 5% de chance d'être due au hasard des données, on considère celle-ci comme *significative* (au seuil indiqué).

Plus concrètement, un test définit un ensemble de valeurs-seuils que doit atteindre la mesure utilisée (χ^2, r, etc.) pour garantir un tel taux d'erreur. Ces valeurs-seuils sont définies par le fait que la valeur ainsi mesurée possède une distribution connue (généralement dépendante des distributions des données initiales), et que l'on connaît donc son comportement sur des données aléatoires.

Si, par le passé, de tels tests se faisaient manuellement (en calculant la valeur visée, et en la comparant avec des tables de valeurs), désormais les logiciels d'analyse statistique donnent directement la probabilité d'erreur, appelée *p-value*, qui est donc généralement indiquée comme résultat quand de telles méthodes sont employées.

Si dans certains contextes les valeurs arbitraires de 5% (ou moins) constituent de véritables seuils à dépasser (pour qu'une publication soit acceptée, qu'une méthode soit validée, etc.), la tendance la plus raisonnable est de considérer la *p-value* comme une indication de la *surprise* apportée par la comparaison de données. L'absence de surprise correspond alors à l'observation que la prise en compte d'une variable n'a aucun effet sur le comportement de l'autre.

Pour tous les tests statistiques, les valeurs seuils ne dépendent pas que de la mesure utilisée, mais également de la complexité et de la taille des données. Par exemple, un coefficient de corrélation de 0,7 mesuré en croisant deux variables numériques n'a pas le même sens s'il est obtenu sur une collection de 10 individus ou sur 1000 : dans le second cas la valeur est bien plus fiable puisqu'elle s'appuie sur un nombre plus grand d'observations, qui garantit de ce fait la stabilité de la dépendance observée. De la même façon, le croisement de deux variables nominales a plus de chance de donner une valeur de χ^2 élevée si chaque variable possède un grand nombre de valeurs possibles (donc si le tableau de contingence contient beaucoup de cellules). La prise en compte de cette caractéristique des données étudiées est effectuée par la notion de nombre de *degrés de liberté* d'un ensemble de données : cette valeur a un impact très important sur la distribution des valeurs auxquelles la mesure sera comparée. De la même façon, le nombre de degrés de liberté d'un ensemble de données est systématiquement calculé et pris en compte par les outils qui effectuent l'estimation [3].

Au final, l'utilisation d'un test statistique demande peu de choses une fois les données recueillies : le choix essentiel consiste à sélectionner le test le plus adéquat pour évaluer la dépendance entre deux variables. Comme on l'a déjà évoqué, le critère principal est lié à la nature (numérique ou nominale) des variables, et à leur distribution. En effet, certains tests ne sont applicables que dans certaines conditions : le cas le plus répandu est l'exigence d'une

3. Cette notion est à mon avis une des plus complexes à saisir dans les utilisations des tests statistiques, puisqu'elle fait appel aux distributions auxquelles sont comparées les valeurs obtenues. On se contentera ici de donner la façon de les calculer pour chaque test.

distribution normale lorsque l'on compare des variables numériques. De tels tests sont dits *paramétriques*. Dans le cas contraire, les tests *non-paramétriques* n'ont pas de telles exigences, mais on généralement tendance à être plus stricts comme on l'a vu pour les coefficients de corrélation (on parle également de tests robustes). Savoir si l'on a affaire ou non à une variable normalement distribuée peut se faire soit par une représentation graphique (histogramme), soit en utilisant des tests spécifiques (comme le test de Shapiro-Wilk).

6.3.2.2 Test du χ^2

Le test du χ^2 est certainement le plus connu et le plus répandu. Il s'applique aux tableaux de contingence résultant du croisement de deux variables nominales. Pour revenir au tableau 6.4 (et aux calculs déjà effectués en section 6.3.1.1) nous pouvons ainsi obtenir une mesure interprétable de l'écart observé entre les répartitions des types de SE dans les différents corpus. Le tableau contient 4 colonnes (types) et 3 lignes (corpus), ce qui donne 6 degrés de liberté ; pour de tels tableaux cette valeur est obtenue en multipliant le nombre de lignes moins un (3-1=2) par le nombre de colonnes moins un (4-1=3) [4]. La valeur critique pour un test du χ^2 à 6 degrés de liberté et un seuil de 5% est de $12,59$. La valeur calculée précédemment est de $118,6$, donc très largement au-delà, ce qui indique que la différence observée entre les répartitions est significative, et que l'on peut donc conclure à l'existence d'une dépendance entre les deux variables. En fait, la *p-value* est extrêmement basse, estimée à $3,2\ 10^{-23}$. Ce nombre microscopique n'a pas de signification particulière à ce stade, on se contente d'ailleurs généralement d'indiquer qu'il est inférieur au seuil le plus bas utilisé pour estimer la significativité, c'est-à-dire $0,001$.

Si le test nous apporte une sorte de garantie quant à l'existence d'une liaison, il ne donne, par contre, pas de détail quand à l'attraction qui peut exister entre les valeurs de chaque variable (quel type de SE est plus/moins fréquent dans quel corpus). Pour ce faire il est nécessaire de calculer les écarts entre les valeurs observées et attendues, et donc de se reporter à un graphique comme le graphe d'association de la figure 6.5.

6.3.2.3 Significativité des coefficients de corrélation

Une fois calculé le coefficient de corrélation (quel qu'il soit parmi les mesures possibles évoquées en 6.3.1.2), il est également possible d'avoir la *p-value* correspondante. Dans ce cas-là le degré de liberté est égal au nombre d'individus considérés moins 2 (donc ici 840-2=838). Quel que soit le coefficient utilisé, la p-value est ici infinitésimale ; c'est généralement le cas lorsque de tels tests sont utilisés pour des quantités de données dépassant quelques dizaines (voir plus loin sur ce point). On peut donc là encore conclure qu'il existe une corrélation positive significative entre le nombre d'indices et le nombre d'items d'une SE.

6.3.2.4 Tests d'analyse de variance

Les tests qui permettent de mesurer l'existence d'une dépendance entre une variable nominale et une variable numérique sont les plus complexes à envisager, et les plus nombreux dans la littérature statistique. Comme indiqué précédemment, la plupart des « classiques »

4. L'intuition derrière cette formule est que, puisque la somme des lignes et des colonnes est connue, on peut déduire certaines valeurs des cellules si l'on connaît celles des autres. Le nombre de degrés de liberté mesure ainsi le nombre de valeurs qui peuvent varier indépendamment.

exigent une distribution normale des variables numériques, ce qui est rarement le cas dans les données que j'ai eu l'occasion d'analyser. Par contre, il existe des tests non-paramétriques qui se basent sur les rangs des valeurs et non les valeurs elles-mêmes, et permettent donc d'aborder les situations que nous avons évoquées plus haut.

La différence entre certains tests provient également du nombre de valeurs de la variable nominale, autrement dit du nombre de séries de valeurs numériques devant être comparées. Le cas le plus simple est celui où la variable nominale n'a que deux valeurs, et le problème d'analyse se ramène donc à la comparaison de deux séries de valeurs numériques. Le test paramétrique le plus classique et sans doute le plus ancien est le test t de Student qui permet de comparer les moyennes de deux séries. Son équivalent non-paramétrique est le test de Mann-Whitney-Wilcoxon qui se base sur les rangs des valeurs.

Lorsque la variable nominale possède plus de deux valeurs, une version plus complexe de ces tests doit être utilisée. Les tests paramétriques utilisables entrent dans la grande famille des ANOVA, comme extension du test de Student. Pour les tests non-paramétriques, c'est le test de Kruskal-Wallis qui doit être utilisé, une extension du test de Mann-Whitney-Wilcoxon.

Pour reprendre l'exemple précédent de la section 6.2.2.2, le test de Kruskal-Wallis confirme également ce qui a été observé, à savoir que la différence est très nettement significative (encore une fois une *p-value* infinitésimale). Comme pour le test du χ^2, il est toutefois nécessaire de faire appel à une représentation graphique pour observer précisément les différences entre les types.

6.3.3 Prendre en compte l'ensemble des informations disponibles : analyses multidimensionnelles

Bien qu'il soit possible de multiplier, comme on l'a vu, les investigations des différentes caractéristiques décrivant un jeu de données, les résultats de ces analyses ne mettent en lumière qu'une partie seulement des phénomènes en jeu. Il existe une famille de techniques qui cherchent spécifiquement à considérer l'ensemble de l'information disponible, et d'en dégager les grandes tendances. On regroupe ces techniques sous le terme d'analyses multivariées. Comme leur nom l'indique, ces méthodes permettent d'aborder l'ensemble des variables, et de ne pas limiter l'analyse en n'en considérant qu'une ou deux à la fois. Si la notion d'analyse multivariée regroupe une grande variété de techniques, je me limiterai ici aux seules approches exploratoires et descriptives (par opposition aux méthodes à visée prédictive, que j'évoque au prochain chapitre) et plus précisément ce que l'on regroupe sous le nom générique d'*analyse factorielle*.

6.3.3.1 Principes des analyses factorielles

On a vu dans ce chapitre et dans le précédent l'intérêt de visualiser globalement la position des individus dans un nuage de points tel qu'on l'obtient par projection en sélectionnant deux variables numériques, que ce soit par le biais des méthodes de visualisation (comme dans la figure 5.13, page 134) ou plus simplement par un diagramme de dispersion (voir figure 6.4). Mais quand on considère l'ensemble des données contenues dans un tableau comme 6.1, on peut voir que chaque individu est en fait positionné dans un espace à n dimensions, n étant le nombre de variables qui décrivent les données. Il n'est donc simplement pas possible d'avoir une vue d'ensemble à moins de procéder à une simplification des données, en réduisant par projection le nombre de dimensions à 2 ou 3. On évoque classiquement comme métaphore

de cette opération le cas d'un dessin sur une feuille (à 2 dimensions donc) d'un objet (à 3 dimensions). Pour être fidèle à la complexité de l'objet, certains angles de vue sont à privilégier : si l'on représente un stylo par exemple en le regardant par la pointe, le dessin va prendre une forme de cercle, qui ne rendra donc pas du tout l'organisation globale de l'objet réel (qui est plutôt longiligne) : il est bien entendu préférable de le dessiner avec un autre angle de vue.

Les méthodes factorielles sont un moyen de calculer le meilleur angle de vue synthétique qui permette de représenter les données dans un espace avec un nombre de dimensions très réduit. Différentes méthodes de calcul sont possibles, mais ces méthodes cherchent toujours à respecter au maximum les propriétés des individus de la collection (notamment les similarités et différences existant entre eux) pour que la représentation finale soit la plus fidèle (que deux individus similaires soient à la fois proches les uns des autres et éloignés de ceux dont ils diffèrent). Il y aura certes une perte importante d'information (que l'on peut estimer), mais un gain inestimable en interprétabilité.

6.3.3.2 Interprétation des facteurs

L'espace de dimension réduite va permettre, en plus d'une représentation des individus, une interprétation des variables : chaque dimension de l'espace d'arrivée est calculée à partir des dimensions (variables) initiales, généralement pas combinaison linéaire. Ces nouvelles dimensions constituent donc en elles-mêmes un résultat très intéressant, puisqu'elles traduisent les combinaisons de variables qui permettent de décrire au mieux les données : les *facteurs*. Ces facteurs possèdent plusieurs propriétés intéressantes :

- chacun d'eux traduit une part de l'information globale initiale contenue dans le tableau de données. Ils sont calculés dans l'ordre décroissant d'information, et chacun d'eux apporte un complément de détails sur les précédents ;
- ils sont orthogonaux, c'est-à-dire qu'ils traduisent des caractéristiques différentes et indépendantes des données ;
- ils sont basés sur les corrélations (ou dépendances) entre les variables initiales, et respectent donc ces informations dont on a vu précédemment qu'elles étaient vitales pour l'analyse.

Les résultats d'une telle analyse sont généralement présentés comme on va le voir par le biais de *cartes factorielles*, c'est-à-dire des graphiques qui résument les dimensions principales des données et positionnent à la fois les individus et les variables dans un nouvel espace facilement interprétable. Mais chaque facteur est également décrit par un ensemble de coefficients associés à chacune des variables initiales : on peut ainsi exprimer pour chaque axe les caractéristiques des individus qu'il traduit, ce qui constitue une information de grande valeur pour la compréhension de l'organisation globale des données.

6.3.3.3 Variantes

Plusieurs techniques différentes existent dans cette gamme de méthodes, qui sont présentées en détails notamment dans Lebart *et al.* (2006). La plus ancienne est l'*analyse en composantes principales* (ACP), qui est celle que j'ai utilisée dans mes travaux. Cette méthode, comme les autres, se base sur un tableau de données individus×variables numériques. Dans le cas où certaines variables sont nominales, elles doivent d'abord être recodées en utilisant des valeurs numériques (par exemple en utilisant les valeurs 1 et 0 pour indiquer

l'appartenance ou non à une classe). Le calcul va se baser sur les différentes corrélations entre les variables, c'est-à-dire sur une matrice qui contient l'ensemble des coefficients de corrélations de tous les couples de variables envisageables. Le résultat de ce calcul va consister en la définition d'un ensemble de nouvelles variables (les composantes principales) qui seront des combinaisons linéaires des variables initiales respectant les principes rappelés plus haut. C'est dans l'espace défini par ces nouvelles variables que les individus vont être positionnés pour leur interprétation. Il existe généralement autant de composantes principales que de variables initiales, mais elles décrivent chacune une part décroissante de l'information initiale (autrement dit de la variation que présentent les données), si bien qu'on peut généralement se limiter à l'observation et l'utilisation de quelques-unes seulement de ces composantes.

Une méthode similaire est appelée l'analyse factorielle exploratoire, dont les résultats sont généralement comparables à ceux de l'ACP. Elle utilise le même type de données (variables numériques ou nominales numérisées), mais nécessite de fixer à l'avance le nombre de facteurs désirés. C'est notamment cette technique qu'a l'habitude d'utiliser Douglas Biber dans ses analyses (voir plus loin).

La tradition française en analyse de données multidimensionnelles, notamment sous la houlette de J.P. Benzécri (Benzécri, 1982), a vu la mise au point de méthodes spécifiques dont la terminologie prête parfois à confusion. Notamment, Benzécri a mis au point la méthode de l'*analyse factorielle des correspondances* (AFC, également appelée analyse des correspondances) qui, elle, ne s'applique initialement qu'à des tableaux de contingence résultant du croisement de deux variables nominales. Cette méthode permet d'interpréter de telles données quand les variables prennent de nombreuses valeurs, comme c'était le cas plus haut lors du croisement entre les types de SE et les corpus. Elle est très utilisée dans les approches lexicométriques, dans lesquelles les textes (également des groupes ou des segments de textes) sont les individus et les mots qu'ils contiennent les variables qui les décrivent. L'AFC permet ainsi d'identifier des dimensions qui permettent de distinguer à la fois des classes lexicales et des classes de textes (voir notamment Lebart et Salem (1994)). Cette méthode a également été étendue pour traiter plus de deux variables nominales, l'analyse des correspondances multiples (ACM).

Plusieurs exemples canoniques d'utilisation de ces méthodes sont déjà bien connus. Hormis les utilisations massives de ces techniques en lexicométrie, c'est sans doute les travaux de Biber qui ont le plus popularisé ce type d'approche en linguistique de corpus. Dès ses premiers travaux (Biber, 1988, 1995), il a développé une méthode basée sur l'utilisation de nombreux traits linguistiques comme autant de mesures de la variation à travers les différents types de textes (en exploitant comme bien d'autres le travail de constitution de corpus partitionnés regroupant différents genres textuels, dans son cas le *LOB corpus*). Sur la base de plusieurs dizaines de traits, calculés pour chaque texte, il a utilisé une analyse factorielle pour identifier les *dimensions* qui expliquent le mieux la variation mesurée par ces traits à travers le corpus. Ces dimensions sont tout d'abord décrites en observant les coefficients associés pour chacune aux traits initiaux : par exemple la deuxième dimension qui ressort de l'analyse factorielle associe des coefficients positifs aux emplois du passé et aux pronoms de troisième personne, et des coefficients négatifs aux verbes au présent et aux adjectifs épithètes. L'analyse de ces coefficients permet d'attribuer une interprétation globale à cette dimension, à savoir que le pôle positif correspond à des textes narratifs, ce qui est par la suite confirmé en observant les coordonnées des textes sur cette dimension. Biber a ainsi ouvert la voie à de nombreux travaux qui se basent sur cette méthodologie, et permettent d'observer le comportement global d'un ensemble de traits à l'échelle d'un corpus.

6.4 Application des méthodes statistiques

Les exemples sur lesquels je me suis appuyé dans la présentation des méthodes ont été sélectionnés pour leur valeur pédagogique, mais pas forcément pour leur intérêt scientifique. Je vais donc présenter ici quelques résultats pertinents de mes travaux qui ont été obtenus par l'application des méthodes statistiques précédentes.

6.4.1 Anatomie des structures énumératives

Le travail sur les structures énumératives du projet Annodis a été très intense en termes d'analyses statistiques, puisqu'il s'agissait de décrire avec plus de précision ce type de structure discursive sur la base de données réelles massives. La plupart des caractéristiques sont décrites dans Ho-Dac *et al.* (2010), mais je vais rappeler quelques-unes des étapes ici.

L'étude du processus d'annotation lui-même a fait intervenir des analyses de données. La première question dans ce type d'annotation manuelle concernait notamment la vérification que les objets étaient bien *annotables*. Pour ce faire, une étape classique dans ce type de travail a consisté à faire annoter les mêmes textes par les trois annotateurs, afin de vérifier l'accord entre leurs décisions. Cette étape a été concluante (environ 70% des structures repérées par un annotateur le sont aussi par un autre, en étant strict sur la comparaison). Par la suite, j'ai vérifié que les principales caractéristiques des structures annotées n'étaient *pas* dépendantes de l'identité de l'annotateur. En utilisant des tests identiques à ceux présentés ci-dessus (test du χ^2 et analyses de variance), j'ai pu vérifier que les annotations suivantes n'étaient pas liées à l'annotateur puisque la *p-value* obtenue dépassait le seuil de 5% : taille des SE, présence d'amorce ou de clôture, nombre d'items. De même, le type de SE n'est pas lié à l'annotateur. La seule liaison significative concernait le nombre d'indices : un des annotateurs notamment était légèrement plus zélé que les deux autres pour identifier des indices nombreux. Ce type de vérification est un bon exemple des résultats positifs que l'on tire d'un test de liaison négatif.

La définition des types de SE a été réalisée et affinée au cours de l'analyse des données. L'idée initiale de celle-ci était de différencier les principaux modes d'insertion d'une structure énumérative dans un texte, en prenant en considération l'articulation de la SE avec la structure logique du document. Une première étape nous avait conduits à définir 5 types : par rapport à l'actuelle typologie le type 4 se séparait initialement en deux sous-catégories. Nous avions distingué les SE qui coïncident avec un paragraphe de celles qui n'en couvrent qu'une sous-partie. Toutefois, aucun test de dépendance entre cette distinction et les caractéristiques formelles des structures n'a permis de mettre en évidence une distinction entre ces deux sous-types. La conséquence a été pour nous de les réunir en un seul.

6.4.2 Difficulté des requêtes en recherche d'information

Dans le cadre du projet ARIEL (évoqué en section 2.3.2), j'ai réalisé avec Josiane Mothe une étude sur les données des campagnes TREC d'évaluation des systèmes de recherche d'information. Ces campagnes proposent chaque année aux participants d'évaluer leur système de recherche d'information sur des requêtes définies pour l'occasion, et ainsi de proposer une comparaison entre les différentes méthodes choisies. Ces campagnes ont l'énorme avantage de mutualiser le coût important lié à l'établissement de jeux de données constitués non seulement des requêtes (textuelles) mais surtout de la liste des documents de la collection-cible qui sont déclarés comme pertinents pour chacune de celles-ci. De plus, les données de toutes

les campagnes sont disponibles, donnant un accès à tous les niveaux de détails : pour chaque système ayant participé à la compétition, on connaît ainsi le score obtenu pour chaque requête (en termes de rappel et de précision).

Notre objectif était d'identifier quelles caractéristiques linguistiques des requêtes pouvaient être associées à des requêtes *difficiles*. La difficulté d'une requête pouvait être approchée dans ce cadre par les scores obtenus par les systèmes qui l'avaient traitée. Nous avons donc utilisé la moyenne de ces scores comme indicateur global (donc une simple variable numérique décrivant chaque requête). Parallèlement à cela, j'ai défini et mesuré un ensemble d'indicateurs sur la base du seul texte des requêtes. Parmi les caractéristiques linguistiques visées par ces mesures, citons : les noms propres, les mots suffixés, la complexité syntaxique mesurée de deux façons distinctes, la polysémie, la ponctuation, etc. Chacun de ces traits se traduisait par une valeur numérique pour chaque requête. Nous avons ensuite effectué des calculs de corrélation entre chacun de ces traits et la difficulté de la requête.

Les résultats sont présentés dans (Mothe et Tanguy, 2005). Plusieurs traits ont atteint le niveau de significativité, indiquant donc qu'ils étaient de bons candidats pour permettre le repérage en amont de requêtes difficiles à traiter. Ces traits sont :

- l'absence de nom propre : un nom propre est un *facilitateur* de requête, comme l'avaient déjà repéré par une étude manuelle Mandl et Womser-Hacker (2002) ;
- la profondeur de l'arbre syntaxique de la requête et la distance entre mots syntaxiquement liés (ces deux caractéristiques étant calculées sur la base des sorties de l'analyseur Syntex) : ces deux mesures sont positivement liées à la difficulté ;
- la présence de mots polysémiques est également corrélée à la difficulté, la polysémie étant ici estimée grossièrement comme étant le nombre de *synsets* de WordNet dans lesquels un mot apparaît.

Une fois validés, ces indices peuvent être utilisés comme prédicteurs de la difficulté, ou encore dans un système d'aiguillage d'une requête vers un mode de traitement plus adapté. La seconde perspective n'a pas encore été atteinte, essentiellement à cause du fait que la difficulté des requêtes est ressentie uniformément par l'ensemble des systèmes. Autrement dit, face à une requête difficile, les méthodes évaluées sont toutes aussi inefficaces les unes que les autres. De plus, il reste difficile de relier précisément ces caractéristiques à un phénomène linguistique local plus précis, ce qui ne permet pas non plus d'envisager une adaptation de la chaîne de traitement (comme les mécanismes de normalisation et de filtrage des termes utilisés pour représenter les documents).

Quoiqu'il en soit, ces indicateurs font désormais partie des indices connus pour être liés à la difficulté d'une requête en RI, comme l'ont synthétisé Carmel et Yom-Tov (2010).

6.4.3 Caractéristiques des consultations médicales

Dans le cadre du projet Intermede, en plus des nouvelles mesures présentées au chapitre précédent, chacune des consultations médicales retranscrites que nous avons étudiées a également été décrite par un ensemble de caractéristiques qui ont fait l'objet d'une analyse statistique.

Les principales caractéristiques suivantes ont été rassemblées :

- nombre total de mots dans la consultation ;
- proportion de mots prononcés par le médecin ;
- part du vocabulaire (classes ouvertes uniquement) prononcé par les deux locuteurs ;
- part du vocabulaire de chaque locuteur qui n'est pas également prononcé par l'autre ;

- nombre de questions posées par chaque locuteur ;
- taux d'énoncés de répétition (dans lesquels le locuteur se contente de répéter quelques-uns des mots de son interlocuteur) prononcés par chaque locuteur ;
- proportion de vocabulaire subjectif prononcé par chaque locuteur ;
- fréquence des modalisations de chaque locuteur ;
- âge et sexe de chaque locuteur ;
- catégorie sociale du patient ;
- type de consultation (caractéristique proposée par les partenaires sociologues sur la base de leurs propres analyses).

Toutes ces caractéristiques ont été croisées deux à deux, en utilisant les tests non-paramétriques présentés précédemment (en fonction de la nature des variables : test du χ^2, ρ de Spearman et test de Kruskal-Wallis). Un ensemble de dépendances entre les variables ont ainsi pu être identifiées, parmi lesquelles on notera :

- les patients prennent plus la parole dans les consultations longues ;
- plus un patient prend la parole, plus il pose de questions ;
- le vocabulaire commun décroît avec le taux de parole du médecin ;
- les patients femmes ont des consultations plus longues (en nombre de mots) et posent plus de questions que les hommes ;
- les médecins femmes prennent moins la parole, ont un vocabulaire partagé plus important et posent plus de questions que les médecins hommes ;
- les patients âgés posent plus de questions ;
- les médecins utilisent moins de modalisateurs avec des patients âgés qu'avec des jeunes, etc.

La liste des dépendances ainsi repérées est encore plus longue, et montre bien la richesse des informations disponibles en observant des caractéristiques linguistiques somme toute assez simples dans ce corpus.

Plutôt que de tirer les conclusions liées à ces résultats, je préfère à ce stade souligner les limites de ces approches bivariées qui se contentent de mesurer la dépendance entre deux variables. Il est en effet clair que, premièrement, la possibilité d'avoir un point de vue global est rendue difficile par la profusion de valeurs qui synthétisent ces comparaisons deux à deux (alors que le rôle des analyses statistiques est principalement d'accéder à cette synthèse), et que deuxièmement cette approche est limitée par la prise en compte de facteurs multiples pouvant « expliquer » les variations mesurées.

Nous avons donc appliqué une analyse factorielle (en l'occurrence une ACP) sur l'ensemble des traits, afin de dégager les dimensions langagières de ces données (Tanguy *et al.*, 2011a). En nous limitant aux seules deux premières dimensions, nous avons pu identifier la représentation synthétique présentée dans la figure 6.6.

Pour analyser ce type de représentation, plusieurs points doivent être pris en considération. Tout d'abord, ne regardons pour l'instant que les seuls éléments identifiés par des flèches sur le schéma : ce sont les seuls traits linguistiques utilisés pour l'analyse. La carte ne présente que les deux premières dimensions issues de l'ACP : la première est représentée sur l'axe des abscisses et la seconde sur l'axe des ordonnées. Chaque trait est donc positionné sur ce graphique, grâce à ses coordonnées dans ce nouvel espace : ces coordonnées indiquent son importance relative par rapport à ces deux axes. Ainsi, les traits les plus pertinents pour un axe sont ceux qui sont situés aux extrémités de la carte (proches du cercle extérieur), par exemple les traits *Parole_P* et *Voc_P* sont situés à l'extrémité droite du graphique, et sont donc très importants pour la première dimension (comme les traits *Parole_M* et *Voc_M* à

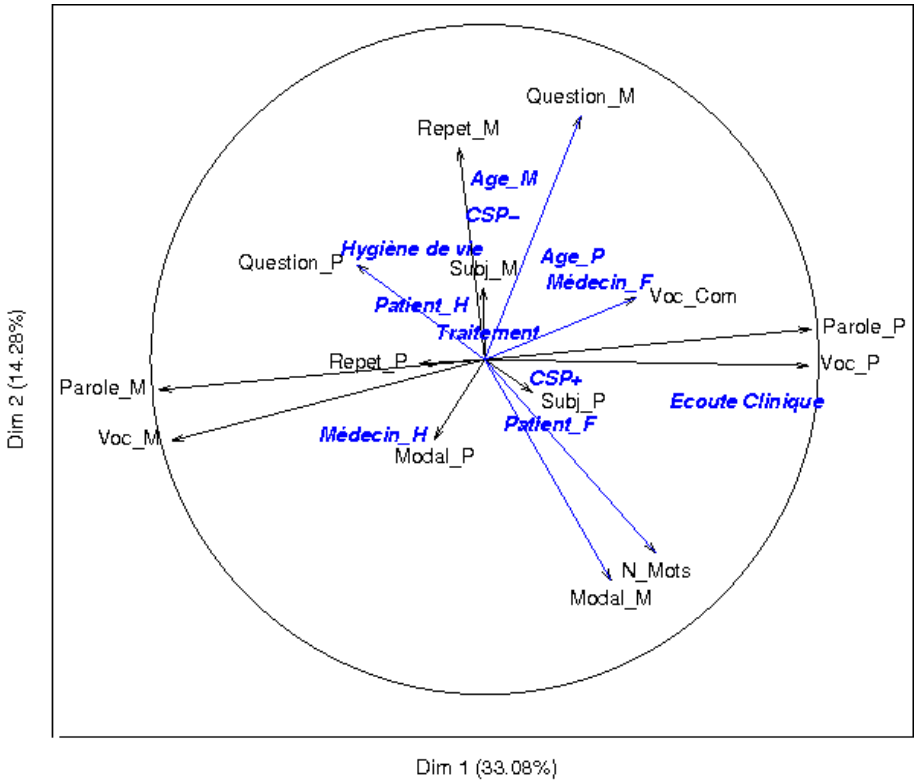

FIGURE 6.6 – Carte factorielle des caractéristiques des consultations médicales

l'extrémité gauche). Ces traits correspondent à la répartition de la parole entre le médecin et le patient : une consultation pour laquelle le médecin est le principal locuteur aura une forte valeur pour le trait *Parole_M* et vice-versa. Les deux autres variables identifient la part de vocabulaire spécifique à chaque locuteur. On peut donc se baser sur ces variables pour interpréter cette première dimension comme étant celle qui oppose les consultations en fonction de leur locuteur principal, celui-ci ayant également tendance à avoir un vocabulaire spécifique (ce second aspect étant une conséquence directe du premier : un patient qui parle peu ne va pas prononcer beaucoup de mots, donc également très peu de mots qui seront aussi utilisés par le médecin). Dans une moindre mesure, on voit que lorsqu'un locuteur mobilise la parole, son interlocuteur a tendance à poser plus de questions (variables *Question_M* et *Question_P*). On voit également (dans une moindre mesure) que les consultations où le patient parle le plus sont également plus longues (la variable *N_Mots* est orientée vers ce pôle). Cette première dimension est comme on l'a dit celle qui « explique » la plus grande part d'information disponible dans les données, ici elle en représente un tiers (33,08%).

La seconde dimension est donc utilisée pour compléter l'information fournie par la première (en quelque sorte une précision de l'approximation correspondante, elle n'ajoute que 14% de l'information totale), amenant le graphique à représenter 47% de la complexité totale. Elle fait intervenir des variables différentes. On peut voir que l'axe vertical oppose (en haut) les consultations où le médecin emploie un vocabulaire subjectif (*Subj_M*), répète les énoncés de son patient (*Repet_M*) et pose des questions (*Question_M*), à celles (en bas) où il utilise des modalisateurs (*Modal_M*), et qui sont sensiblement plus longues. On peut voir dans cette

seconde dimension une différence de comportement énonciatif du médecin.

Les consultations analysées étaient également décrites par un ensemble de variables *externes*, correspondant notamment aux caractéristiques du patient et du médecin. Ces variables, représentées en gras sur le graphique, ont été considérées comme illustratives, c'est-à-dire qu'elles ne sont pas prises en compte dans le calcul des dimensions (qui sont donc uniquement basées sur les caractéristiques linguistiques mesurées), mais sont repérables dans le nouvel espace organisé par celles-ci. On peut donc enrichir la description de ces deux dimensions par leur relation avec ces variables : on voit ainsi que les médecins qui laissent le plus parler leurs patients sont plutôt des femmes, et que le type de consultation correspond au modèle de *l'écoute clinique*. Sur la seconde dimension, le premier pôle est plutôt du fait de médecins âgés (*Age_M*) s'adressant à des patients âgés (*Age_P*).

Comme on le voit, ce type d'analyse permet donc d'observer un ensemble de régularités assez facilement interprétables, bien plus que ne le permettait la somme des corrélations mesurées entre ces mêmes variables. Elle permet également surtout d'envisager des comportements de groupes de traits (comme ceux qui organisent chaque dimension de l'analyse factorielle), là où l'analyse bivariée ne donnait que des relations isolées.

Il est par contre nécessaire de rester prudent vis-à-vis de ce type de représentation. Premièrement, l'ensemble du graphique avec les deux dimensions ne représente qu'un peu moins de la moitié de l'information totale. De plus, certains rapprochements entre traits effectués par la représentation graphique peuvent à tort être interprétés comme des corrélations significatives, ce qui n'est pas nécessairement le cas, la projection sur le plan des deux seules premières dimensions entraînant des effets d'optique qui peuvent tout simplement induire en erreur.

Il existe d'autres résultats produits par ce type d'analyse et qui ne sont pas exploités. Notamment, en plus de la carte factorielle qui positionne les différentes variables sur un graphique, il est possible d'y placer aussi les individus. Autrement dit, il est possible de visualiser, tout aussi simplement, quelles sont les consultations qui correspondent à chacun des profils décrits par les deux axes, et ainsi de tendre vers une typologie effective des consultations médicales. Ce n'était toutefois pas le but de notre étude.

6.4.4 Complexité syntaxique : identification de dimensions

De façon peut-être plus anecdotique, mais ayant permis d'identifier des phénomènes intéressants, je me suis penché sur la notion de complexité syntaxique dans le cadre du mémoire de master de Nikola Tulechki ayant donné lieu à une publication (Tanguy et Tulechki, 2009). Cette étude voulait étudier à travers une étude sur des données relativement volumineuses les différentes façons dont est mesurée, à travers un panel d'applications, la complexité syntaxique. Que ce soit pour estimer la lisibilité d'un texte, pour définir un langage contrôlé ou (comme nous l'avions proposé dans Mothe et Tanguy (2005)) pour prédire la difficulté d'une requête en recherche d'information, plusieurs mesures ont été proposées. La plupart de ces mesures sont calculables automatiquement sur la base des sorties d'un analyseur syntaxique, si bien que nous avons construit un tableau de données qui, pour 130 000 phrases extraites de différents types de textes, donne la valeur correspondant à chacun des 52 traits sélectionnés.

L'utilisation d'une analyse multidimensionnelle visait là aussi à identifier les principales dimensions couvertes par ces différentes définitions de la complexité. En utilisant une ACP, nous avons pu observer quels étaient les traits qui avaient un fonctionnement similaire, mais surtout ceux qui mesuraient des aspects différents de la complexité. Ce deuxième aspect est

très important, et c'est celui qui est une conséquence de l'orthogonalité des composantes principales produites par une ACP. Nous avons proposé l'interprétation suivante pour les trois premiers facteurs :

1. Longueur de la phrase (26%) : ce facteur est le moins intéressant de tous, car le plus trivial. Cette dimension permet néanmoins de confirmer que la totalité des mesures de lisibilité proposée en psycholinguistique, malgré leur sophistication progressive, se résument à cette approche naïve de la complexité. De plus, le fait que cette dimension soit clairement ressortie en premier permet d'aborder les autres aspects, qui sont eux indépendants de la longueur.
2. Nature verbale *versus* nominale de la phrase (10%) : la complexité d'une phrase peut en effet provenir de la présence de syntagmes nominaux complexes, ou bien de constructions verbales (subordonnées, conjonctions verbales).
3. Connexité syntaxique (8%) : cette dimension oppose des phrases (indépendamment de leur longueur et leur nature nominale/verbale) dont les différents constituants sont effectivement reliés par des relations syntaxiques établies, à celles qui utilisent des constructions plus déstructurantes (appositions, groupes circonstants, parenthèses, etc.).

Une fois établies, ces dimensions permettent, comme l'a fait Biber, d'observer les positions relatives de différents types ou genres de textes, et d'affiner ainsi leur définition sur ce plan.

6.4.5 Comparaison des résultats en recherche d'information : catégorisation des systèmes

Un troisième exemple d'utilisation de méthodes d'analyse factorielle provient de mon travail avec Josiane Mothe dans le cadre du projet ARIEL qui visait à la définition d'un système de recherche d'information adaptatif, i.e. qui modifierait ses traitements en fonction des requêtes. L'essentiel de notre travail a porté sur l'étude de la variation dans les performances des systèmes de RI qui ont participé aux campagnes d'évaluation TREC (Chrisment *et al.*, 2005; Mothe et Tanguy, 2007).

Par exemple, pour la campagne TREC Novelty 2002, nous avons étudié les résultats des 42 systèmes participants en considérant leur score de rappel pour les 50 requêtes qui constituaient le jeu d'évaluation : le tableau de données utilisé comprend donc les 42×50 valeurs de rappel.

L'utilisation d'une ACP a tout d'abord permis d'observer les dimensions de la variation des performances des systèmes en considérant les scores pour chaque requête comme autant de variables. Il ne s'agissait pas ici d'identifier la nature des dimensions qui séparaient les différents systèmes, mais plutôt d'identifier différents groupes parmi ceux-ci, afin d'en capter la complémentarité : si certains types de systèmes ont des comportements distincts en fonction des requêtes, ils seront identifiables par une ACP. Bien que la variation globale des performances des systèmes soit globalement faible (à part quelques systèmes sous-performants), il a en effet été possible d'identifier deux groupes qui se positionnaient de façon orthogonale sur la carte factorielle, comme indiqué dans la figure 6.7.

Ces deux groupes de systèmes montraient effectivement des comportements différents sur certaines requêtes, et leur utilisation combinée était donc une piste à suivre pour envisager une amélioration globale des performances. Malheureusement, un élément essentiel de l'analyse consistait en la détection des types de requêtes pour lesquelles cette variation était constatée : nous n'avons pas pu mettre en évidence de lien entre celles-ci et les caractéristiques linguistiques utilisées avec succès pour identifier les requêtes difficiles (voir section 6.4.2).

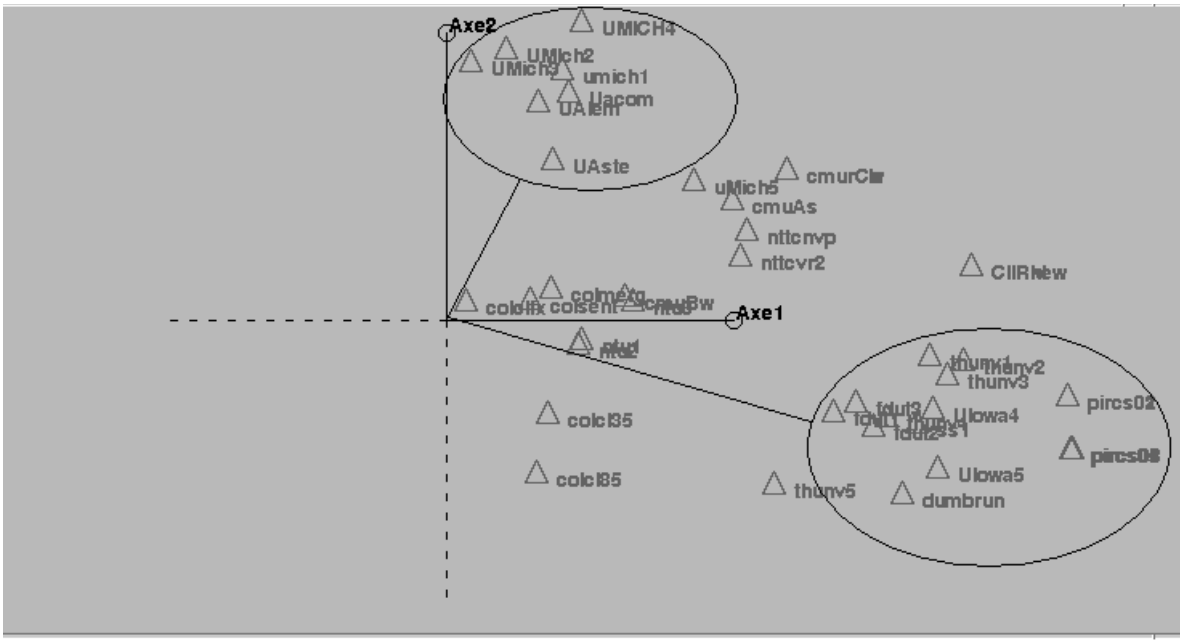

FIGURE 6.7 – Carte factorielle des différents systèmes de la campagne TREC Novelty 2002

Malgré ce résultat décevant, et le fait que les données analysées n'ont qu'un très lointain rapport avec le matériau langagier, on voit bien par cet exemple la plasticité des utilisations de ce type de méthode. Ce dernier aspect montre également qu'une approche de l'analyse des données consiste effectivement à mettre au jour des classes d'individus en fonction de leurs descriptions par un ensemble de variables. Ce point sera détaillé au chapitre suivant, même si les techniques utilisées entrent de fait dans la grande famille des méthodes statistiques multivariées.

6.5 Complexité, limites et pertinence des analyses statistiques

Pour clore ce chapitre sur l'utilisation des méthodes statistiques, je souhaite faire une sorte de bilan de mes expériences dans ce domaine, et surtout soulever un ensemble de points méthodologiques qui peuvent donner un petit sentiment de malaise face à ce genre d'approche des données langagières.

6.5.1 Des approches doxiques mais utiles

J'ai insisté au début de ce chapitre sur la pression sociale que je perçois comme étant une des motivations principales de l'utilisation croissante des méthodes statistiques en linguistique de corpus. Je suis néanmoins persuadé de leur utilité scientifique dans nombre de situations, essentiellement exploratoires.

Dans la continuité des différents types de représentations visuelles présentées au chapitre précédent, on a vu ici que la statistique descriptive fait, elle aussi, un usage intensif des graphiques pour aborder les données langagières. Le rôle de ces derniers est double : ils

permettent bien sûr le même point de vue global sur les caractéristiques, et permettent à la fois de dégager rapidement les grandes tendances et les cas particuliers, mais ils sont aussi des guides pour les méthodes numériques à employer pour des analyses plus fines. À la différence des techniques de visualisations plus innovantes, les graphiques de la statistique sont par contre nettement plus normés et systématiques (histogrammes, diagrammes de dispersion, cartes factorielles, etc.).

D'un point de vue technique, les différentes méthodes de calcul et de représentation graphique évoquées dans ce chapitre sont d'une simplicité remarquable à utiliser une fois acquis les principes de base de la manipulation des données, par le biais de logiciel dédiés comme R. Toutes les mesures sont en effet aisément accessibles par des fonctions prédéfinies et prêtes à l'emploi, avec un paramétrage par défaut. Si cette facilité est bien adaptée à l'application d'un protocole bien défini (comme c'est le cas pour des questions bien délimitées par une discipline, comme les analyses de variance en psycholinguistique), il réside toutefois un danger dans l'opacité de fait qu'elle induit, en cachant les détails du calcul et les choix de simplification parfois effectués. Comme pour d'autres outils techniques en linguistique (on l'a vu pour l'interrogation de corpus), il est donc important de veiller à un compromis entre la facilité et la disponibilité, et la compréhension plus fine des mécanismes en jeu. Si j'ai préféré ne pas étaler de formules mathématiques dans ma présentation, il est clair que celles-ci doivent être abordées tôt ou tard.

Le rôle principal que j'ai tendance à attribuer à ces méthodes est celui d'un tamis quand on se retrouve face à un nombre important de caractéristiques dont il s'agit de démêler les interrelations. Quelle que soit la fiabilité mathématique avérée des tests et des autres mesures, ils permettent d'effectuer ce nécessaire tri, en donnant des pistes pour une investigation plus poussée, et d'éliminer rapidement des caractéristiques peu pertinentes ou redondantes. Cette possibilité de filtrer les données est notamment très utile en cours de traitement, lorsque l'on envisage différentes mesures pour un même phénomène (notamment les différentes mesures liées au vocabulaire, qui peuvent prendre en compte différentes sélections en fonction des catégories grammaticales prises en compte, des seuils de fréquence, etc.) : la mesure d'une forte corrélation entre ces alternatives permet de relativiser l'intérêt de multiplier les paramètres, et la liaison avec une autre variable correspondant a priori à d'autres phénomènes est une piste pour privilégier telle ou telle approche.

6.5.2 Multiplicité des méthodes et des avertissements

J'en arrive maintenant à l'exposé d'un ensemble de questions et de doutes ressentis au fil de mon apprentissage des méthodes statistiques. Pour des techniques qui me semblaient éprouvées, la plupart datant de plusieurs dizaines d'années et utilisées à travers le monde, les époques et les disciplines, je suis étonné du manque de consensus exprimé par les statisticiens sur leur réelle validité mathématique. À cela s'ajoutent les problèmes d'adéquation entre le type d'objet étudié (et notamment les caractéristiques de sa distribution) et des méthodes utilisées (on l'a vu pour les méthodes paramétriques, dont les précautions d'emploi ne sont au final que rarement respectées). On ne compte plus les publications qui remettent en cause les résultats obtenus dans une étude antérieure au vu des techniques statistiques employées. La linguistique de corpus, contrairement à la psycholinguistique, est peut-être spécialement bien dotée de ce côté-là, et les méthodes quantitatives semblent en permanence remises en cause sans avoir le temps de devenir des standards. Espérons que cela ne constitue que le passage obligé d'une discipline en train de se former, et d'une méthodologie en train de se stabiliser.

Ce doute quant au bon usage des statistiques est également valable pour le TAL, comme le montre cette position de Dunning (1993) qui ne mâche guère ses mots :

> « *There has been a recent trend back towards the statistical analysis of text. This trend has resulted in a number of researchers doing good work in information retrieval and natural language processing in general. Unfortunately much of their work has been characterized by a cavalier approach to the statistical issues raised by the results. The approaches taken by such researchers can be divided into three rough categories.*
>
> 1. *Collect enormous volumes of text in order to make straightforward, statistically based measures work well.*
> 2. *Do simple-minded statistical analysis on relatively small volumes of text and either 'correct empirically' for the error or ignore the issue.*
> 3. *Perform no statistical analysis whatsoever.* »

Hormis des constatations à large spectre comme celle-ci, le champ sur lequel la plupart des discussions sont focalisées est sans doute celui qui est aussi le plus fréquenté par des utilisations statistiques, celui de l'étude des fréquences lexicales dans les textes. La linguistique de corpus de l'école anglaise fait en effet un usage direct et systématique des outils statistiques pour essentiellement deux tâches :

- comparer la fréquence d'un mot (ou d'une structure syntaxique) entre deux corpus (et proposer une explication pour la différence observée) ;
- rechercher les cooccurrences *pertinentes* d'un mot dans un corpus, éventuellement en comparant également les différences obtenues sur plusieurs corpus.

Ces deux modes d'approches sont massivement utilisés et implémentés pour un usage direct dans les outils de base que sont les concordanciers. Les différentes façons de comparer des fréquences dans des corpus de taille différentes, ou de mesurer des associations autour de mots-pivots de fréquences différentes sont légion : rapport de vraisemblance, t-score, z-score, information mutuelle, etc., un inventaire précieux étant notamment disponible dans la bible de Manning et Schütze (1999). Toutes ces mesures sont basées sur l'estimation de l'effet de *surprise* par une comparaison avec une distribution aléatoire. Par exemple, elles permettent de voir si les fréquences relatives d'un mot sont identiques dans deux corpus contenant des textes de type différents, ou si la présence d'un mot dans le voisinage d'un pivot est la même que dans le reste du corpus.

Mais les débats autour de la validité de ces mesures, ou leurs conditions d'application font rage : si Manning et Schütze se contentent d'émettre les recommandations habituelles sur les conditions de normalité ou mieux, sur le type de phénomène que chacune d'elles est à-même de capter, certains auteurs sont plus virulents. C'est le cas de Kilgarriff (2005), dont la conclusion est assez radicale :

> « *Language is non-random and hence, when we look at linguistic phenomena in corpora, the null hypothesis will never be true. Moreover, where there is enough data, we shall (almost) always be able to establish that it is not true. In corpus studies, we frequently do have enough data, so the fact that a relation between two phenomena is demonstrably nonrandom, does not support the inference that it is not arbitrary. Hypothesis testing is rarely useful for distinguishing associated from non-associated pairs of phenomena in large corpora. Where used, it has often led to unhelpful or misleading results.* »

Kilgarriff met le doigt sur le problème de la taille des données utilisées pour ce type d'étude : comme la taille des corpus croît rapidement, les valeurs des effectifs utilisés (notamment dans les tests du χ^2) entraînent mécaniquement un dépassement du seuil de significativité. Rappelons que la plupart des méthodes statistiques ont justement été mises au point pour pouvoir inférer des comportements au niveau d'une population en ne travaillant que sur des échantillons de celle-ci [5]. Dans certains manuels de statistique on parle de *grands échantillons* quand la collection contient plus de 30 individus, les tables des valeurs-seuils ne vont au mieux que jusqu'à 100 degrés de libertés (pour le coefficient r de Pearson) : que faire alors des effectifs obtenus par des calculs automatiques sur des corpus de plusieurs centaines de millions de mots, quand un mot apparaît plusieurs milliers de fois, et que la p-value obtenue donne des valeurs vertigineusement basses (comme on l'a vu dans les exemples ci-dessus) ? Les propositions de Kilgarriff face à cet état de fait l'ont conduit à ne plus se baser sur ces méthodes et leurs seuils, et donc à considérer que les valeurs obtenues ne sont pas des estimations obtenues par un processus d'échantillonnage, mais une description précise de l'ensemble de la population. Pour les problèmes typiques qu'il aborde, et qui concernent essentiellement le classement des collocations, on peut effectivement se passer de tels seuils, mais pour d'autres cas il est cependant nécessaire de prendre de telles décisions (aussi arbitraires soient-elles).

Dans un autre domaine que j'ai pu aborder, celui de la recherche d'information, les débats sont tout aussi vifs. Dans le cadre des campagnes d'évaluation à la TREC, les tests statistiques sont utilisés pour comparer les mérites de deux systèmes de RI sur une collection de référence (Hull, 1993). Il s'agit donc concrètement d'utiliser des méthodes reconnues pour valider qu'une différence (parfois minime) entre deux séries de mesures (rappel ou précision) correspond à une véritable variation de performance généralisable en dehors du banc de test. De nombreuses publications traitent donc de la légitimité et du mérite relatif des différents tests statistiques utilisés par la communauté, dont on peut espérer qu'ils convergent dans un avenir proche car ils traitent d'un ensemble homogène de situations d'évaluation (Sanderson, 2010). Plusieurs propositions intéressantes dans ce cadre vont dans le sens de l'abandon des tests classiques comme ceux que j'ai présentés ici, en insistant notamment sur l'évolution des puissances de calcul depuis l'époque de leur invention. Smucker *et al.* (2007) plaident ainsi pour l'utilisation de méthodes de test par randomisation (c'est-à-dire en répétant les mesures sur un très grand nombre d'échantillons aléatoires extraits de la collection étudiée), plus simples et moins exigeantes sur le plan théorique que les tests classiques et qui, si elles nécessitent de nombreux calculs, sont à la porté des machines contemporaines [6]. Il n'est cependant pas sûr que cela change effectivement les pratiques des chercheurs, ni que cela fasse au final avancer le débat sur les modalités d'évaluation plus globales imposées par ce type de campagnes (Jäverlin, 2009).

6.5.3 Vers des méthodes toujours plus complexes

Face à ces difficultés reconnues pour les mesures de fréquences en linguistique, plusieurs solutions sont proposées. Gries (2005) propose de compléter ces calculs par des mesures qui

5. Par exemple, les travaux fondateurs de Student sur le t-test ont été réalisés alors que ce dernier (dont Student est le nom de plume) travaillait sur la mise au point d'une méthode de vérification de la qualité de la bière dans les brasseries Guinness, en ne testant bien entendu qu'une infime partie de la production.

6. Le test de Wilcoxon par exemple, doit sa popularité à l'extrême simplicité des calculs nécessaires, qui peuvent aisément se faire à la main. L'idée des nouvelles méthodes est de proposer un changement de point de vue sur la question au vu de l'outillage moderne.

permettent la prise en compte de la taille (comme on l'a vu plus haut avec le coefficient V de Cramer), mais sans véritable indication de ce qui doit ou ne doit pas être retenu au final (malgré la grande pédagogie de Gries dans ses nombreuses publications adressées aux linguistes de corpus, peu de décisions tranchées sont proposées dans ses approches).

De façon plus ambitieuse, d'autres proposent d'utiliser des mesures qui prennent en compte la distribution connue des mots dans les textes. On sait notamment depuis les travaux de Zipf que la distribution des mots dans un texte a un comportement à peu près stable, et que (grossièrement) une part importante du vocabulaire d'un texte est constitué de mots qui n'apparaissent que très rarement, alors qu'un petit ensemble de formes seront massivement répétées (ce type de modèle a pour nom LNRE pour *Large Number of Rare Events*). Une distribution normale indiquerait au contraire qu'une majorité de mots ont une fréquence moyenne, avec quelques mots aux comportements plus extrêmes (très rares ou très fréquents, mais de façon symétrique). Dès lors, il est possible d'envisager de construire des modèles statistiques correspondant aux phénomènes précis que la linguistique de corpus étudie plutôt que d'utiliser des outils notoirement inadaptés. Malheureusement, le bagage mathématique nécessaire à la compréhension, la modélisation et l'utilisation de ces méthodes est largement au-delà du bagage statistique de base (au moins du mien), comme le montre la complexité des travaux des spécialistes de ces questions comme Baayen (2001). De plus, ces approches sont, rappelons-le, quasi-exclusivement destinées à l'étude de la répartition du vocabulaire dans les corpus : si la distribution LNRE s'applique également à certains types de marques lexico-syntaxiques, ce n'est pas le cas de tous les phénomènes étudiés en corpus, ni des caractéristiques externes utilisées pour qualifier les productions langagières (l'âge des patients dans Intermede par exemple).

Deux autres points me semblent également poser des difficultés méthodologiques importantes pour le type d'étude que j'ai évoquées.

Le premier concerne la nécessité de « faire rentrer » les données étudiées dans le moule des fameux tableaux individus×variables évoqués au début du chapitre. Comme cela transparaît clairement pour l'étude des structures énumératives, les données langagières sont bien souvent des compositions complexes organisant des objets de différents niveaux. Ce ne sont en effet ni des mots ni des phrases que l'on examine, mais des zones de textes de taille arbitraire et variable, elles-mêmes composées de segments, qui entretiennent entre eux des relations comme la position relative, et qui sont placées dans un texte lui-même constitutif d'un corpus. Dès lors, une approche statistique qui exige d'identifier un type d'individu sur lequel vont se porter les efforts et les questions va nécessiter une simplification importante de l'information prise en compte. C'est ainsi que nous avons réduit la notion d'indices des SE à un simple comptage, éventuellement après un typage sommaire. Ces approximations sont nécessaires pour dégrossir l'analyse, mais il serait dommage de les garder plus en aval au seul motif que les outils d'analyse ne peuvent les prendre en compte.

Le second concerne les études multivariées dont nous n'avons évoqué qu'une seule famille de méthodes (les analyses factorielles). Face à la question de l'interaction de plus de deux variables, une importante gamme de techniques statistiques concerne la notion de régression. Exprimée de façon simplificatrice, une régression est la recherche d'un modèle qui permet de relier une variable (appelée variable dépendante) à une ou plusieurs autres variables (appelées variables indépendantes ou prédictives). Par exemple, l'observation d'une corrélation linéaire importante entre deux variables numériques permet d'en déduire une formule qui relie les deux séries de valeurs. On entre alors dans la problématique des statistiques prédictives : un tel modèle permet en effet, sur la base de données observées, de calculer la valeur probable de

la variable dépendante pour un individu dont on ne connaît que les variables prédictives (par exemple, on peut estimer grossièrement d'après les données observées dans Annodis le nombre d'indices qui marque une SE dont on ne connaît que le nombre d'items). Bien que cet objectif de prédiction ne soit pas celui qui a été abordé ici (mais je l'aborde dans la dernière partie du mémoire), l'établissement d'un *bon* modèle est une façon d'approfondir la connaissance des objets qu'il décrit. Les techniques de régression qui font intervenir plusieurs variables prédictives peuvent notamment mesurer l'importance relative de plusieurs paramètres dont on a pu mesurer (par des techniques bivariées) qu'ils étaient liés à une variable ciblée. On a vu par exemple pour les données d'Intermede que la longueur de la consultation était liée à l'âge et au sexe des protagonistes, mais les participations relatives de ceux-ci ne sont pas accessibles sans prendre en compte leur contribution conjointe. Il existe notamment des effets qui font intervenir certaines combinaisons de facteurs (on a pu observer par exemple quelques variations significatives en regardant si le médecin et le patient étaient du même sexe). Les régressions multiples sont des techniques qui peuvent donc approfondir la mise au jour de tels phénomènes, mais leur complexité est bien plus grande que celle des approches statistiques de base vues ici.

L'enjeu est pourtant très important pour un ensemble de questions, par exemple dans l'étude de la signalisation des phénomènes discursifs. Comme on le verra dans la partie suivante, si certaines caractéristiques contextuelles d'un segment de texte ne sont pas suffisantes pour signaler la nature ou la fonction de celui-ci, il est possible qu'elles aient un tel pouvoir lorsqu'elles sont combinées, comme on l'a vu brièvement pour les différents indices des SE.

Comme je viens de le dire, ces techniques statistiques complexes (régression linéaire multiple, régression logistique, etc.) ont été développées essentiellement pour leur rôle de prédiction. Il se trouve qu'en TAL, comme dans d'autres domaines faisant intervenir le traitement de données massives, cette préoccupation rejoint celle d'un champ dont l'ampleur n'a également fait que croître au fil des ans, l'apprentissage automatique et la fouille de données. Que ce soit par pression sociale, par curiosité intellectuelle ou par facilité, j'ai donc préféré aborder certaines des questions de ce type en faisant appel à ce type d'approche, comme je vais maintenant le présenter dans la dernière partie de ce mémoire.

Quatrième partie

Exploiter la complexité des données : fouille et apprentissage automatique

C'est maintenant un fait acquis que la grande majorité des développements actuels en TAL concernent des méthodes basées sur des données, qui mobilisent massivement les techniques d'apprentissage automatique pour concevoir un système finalisé. Que ce soit pour la traduction automatique, l'étiquetage de textes, le résumé automatique, la classification de documents etc., ce ne sont plus des règles issues de descriptions linguistiques qui constituent le cœur mécanique de ces systèmes. Au contraire, les informations nécessaires sont extraites automatiquement de grandes masses de données, généralement annotées manuellement pour servir de base d'apprentissage.

Dans le même ordre d'idée, et en utilisant des méthodes similaires, les techniques de fouille ont été développées pour assister l'analyse de grandes quantités de données, et pour faire émerger des régularités dans les caractéristiques de celles-ci. Elles sont elles aussi de plus en plus régulièrement appliquées aux données langagières annotées (ou non) comme mode d'exploration.

Ce sont ces deux aspects intimement liés que je vais aborder dans cette dernière partie.

Je commencerai dans le chapitre 7 par présenter ces techniques, en montrant comment j'ai pu utiliser certaines d'entre elles avec des objectifs de deux types :

- l'analyse de données langagières annotées, dans le même esprit que les explorations visuelles et les analyses statistiques présentées au chapitre précédent ;
- le développement d'outils de TAL avec un objectif plus applicatif.

J'y montrerai notamment comment ces techniques s'intègrent dans le travail d'investigation, et les avantages que l'on peut en tirer.

Dans le chapitre 8, je propose d'élargir la problématique et la réflexion autour des changements qu'entraînent ces nouvelles techniques. Désormais accompli et incontournable, ce changement de paradigme arrive en effet à l'heure d'un bilan plus profond et contrasté, que j'essaie d'aborder en insistant sur les problèmes qu'il pose aux facettes plus linguistiques du TAL auxquels je suis sensibilisé. Il fait écho aux réflexions de mes collègues proches, dont Cécile Fabre (2010) avec qui je partage ces interrogations disciplinaires mais aussi plus concrètes. Dans cette situation souvent inconfortable de rupture au sein de la communauté du TAL, j'essaierai de dégager les principaux points de ruptures, et surtout la façon de les aborder. Je terminerai par l'exposé de travaux récents qui, s'ils s'inscrivent plus facilement dans le nouveau paysage du TAL, cherchent toutefois à questionner autrement les techniques et les données en affrontant l'opacité des premières et la distance avec les secondes qui semblent s'accroître avec ces nouvelles approches quantitatives.

Chapitre 7

Utilisation des méthodes de fouille de données et d'apprentissage automatique

Je tiens ici à préciser que je ne suis pas plus un spécialiste de l'apprentissage automatique que des méthodes statistiques, et que je me situe simplement comme un utilisateur (éclairé), là encore cherchant à tirer parti du courant dominant en l'appliquant à mes pratiques et à celles de mon entourage disciplinaire proche, en cherchant toutefois à en préserver les spécificités. Les techniques d'apprentissage automatique « classiques » font néanmoins partie de ma culture initiale, puisqu'elles sont directement issues de l'intelligence artificielle, et que leur utilisation était déjà très répandue lorsque j'étais étudiant en informatique. Cette connaissance (hors du domaine du TAL) m'a permis de pouvoir suivre plus facilement les évolutions dans leurs applications, du moins jusqu'à un certain point, même si je ne les ai utilisées que récemment et si elles ont bien entendu beaucoup évolué depuis l'époque de ma formation en IA.

Dans ce chapitre, je vais tout d'abord présenter les grands principes des méthodes par apprentissage, en expliquant leur fonctionnement et leurs conditions d'emploi. Ensuite, je présenterai les utilisations que j'ai pu en faire. J'ai séparé celles-ci en deux catégories, en fonction des objectifs visés. Les premiers travaux présentés sont ceux qui, dans la continuité des questions exploratoires autour des données langagières annotées, cherchent à dégager les principales caractéristiques des objets linguistiques étudiés. Les seconds correspondent à des objectifs plus appliqués, et utilisent les méthodes par apprentissage pour répondre à un besoin relevant de la classification ou de l'annotation automatique.

7.1 Principales notions et méthodes

Il n'est pas toujours facile de distinguer l'apprentissage automatique (*machine learning*), la fouille de données (*data mining*), sa spécialisation en fouille de texte (*text mining*) et l'analyse statistique. Les trois disciplines s'attaquent aux mêmes types de données et utilisent des principes et des algorithmes similaires quand ils ne sont pas identiques. Toutes trois se penchent sur des ensembles de données, et tirent un bénéfice direct d'un accroissement de la taille de celles-ci. Elles se basent sur le repérage de *régularités* dans ces données, tout en mesurant la stabilité et la fiabilité de celles-ci.

Les techniques statistiques sont simplement le fondement sur lequel se basent les deux

autres disciplines, plus récentes. La fouille de données et l'apprentissage automatique concernent l'utilisation et le développement d'outils informatiques qui automatisent des procédures de traitement spécifiques. La fouille de données implique nécessairement l'utilisation de très grandes quantités de données, et n'a de fait émergé comme discipline qu'avec leur disponibilité croissante (les fameux *entrepôts de données* qui accumulent les informations produites par toute activité institutionnelle).

On a vu que les analyses statistiques permettaient soit une meilleure compréhension des données décrites (en présentant des synthèses, et en identifiant des associations entre variables), soit la construction de modèles capables de prédire la valeur probable d'une variable en fonction des autres.

L'apprentissage automatique vise à exploiter un ensemble de données pour pouvoir traiter efficacement de nouvelles données. Pour ce faire, les systèmes d'apprentissage automatique se basent sur le repérage de régularités dans les données analysées, qui sont ensuite formalisées de façon à être exploitées dans un processus de décision.

La fouille de données vise à la découverte de nouvelles informations sur la base d'une collection de données. Ces connaissances proviennent du repérage de régularités dans les données, et sont prévues pour être présentées de façon intelligible. Elles peuvent éventuellement servir à prendre des décisions pour traiter de nouvelles données similaires, auquel cas on rejoint alors le domaine de l'apprentissage.

Les techniques d'apprentissage ne permettent cependant pas toutes la découverte de ces régularités sous une forme intelligible : seules les techniques symboliques proposent des modèles sous la forme de règles logiques (si A alors B) ou de formalismes équivalents. Les autres techniques utilisent des modèles numériques et probabilistes (comme les réseaux de neurones, les systèmes bayésiens ou les SVM) qui sont par nature opaques et dont la seule utilisation possible est l'application à de nouvelles données pour calculer une valeur ou déduire l'appartenance à une classe.

Je résumerai donc la différence entre ces méthodes de la façon suivante : l'apprentissage automatique vise à construire un programme qui s'applique à un problème donné (de classification, ou d'estimation d'une valeur) en se basant sur des données déjà traitées qui sont prises comme base d'apprentissage.

7.1.1 Apprentissage supervisé

L'objet le plus classique de l'apprentissage automatique consiste, tout comme pour les statistiques descriptives, en un ensemble d'objets distincts décrits par un ensemble de variables. Dans le cas de l'apprentissage supervisé, une de ces variables est sélectionnée comme étant la cible du processus : c'est cette variable qui devra être prédite en se basant sur les autres.

7.1.1.1 Principes

Le cas le plus simple et le plus courant est l'attribution d'une catégorie prédéfinie à chaque élément d'un ensemble d'individus : c'est ce que fait l'étiquetage (morpho-syntaxique) pour les mots d'un texte et la classification de texte (thématique ou autre) pour les textes d'une collection. Les objets à traiter sont représentés par un ensemble de descripteurs (variables), généralement obtenus par une analyse simple et automatique (forme, position et environnement immédiat des mots, termes contenus dans un texte, etc.). Le processus d'apprentissage automatique se déroule ainsi :

1. sélection et catégorisation (généralement manuelle) d'un ensemble de données qui vont constituer le corpus d'apprentissage ;
2. définition et projection automatique d'un ensemble de descripteurs sur le corpus d'apprentissage ;
3. analyse de ces données en recherchant des régularités permettant d'associer la catégorie ciblée aux descripteurs (phase d'apprentissage) ;
4. formalisation de ces régularités par un ensemble de règles ou une formule numérique (construction du modèle) ;
5. application de ces règles ou de cette formule à des données non catégorisées (phase d'exploitation).

L'intervention humaine dans ce processus se limite donc (au mieux) à la constitution des données d'apprentissage, tâche reconnue comme étant longue et fastidieuse puisque nécessitant généralement un travail manuel d'analyse, et devant atteindre un volume minimal et une représentativité suffisante pour garantir la validité du modèle. Bien sûr, le paramétrage du système et son évaluation sont des phases essentielles, mais dont l'automatisation peut être importante : les données déjà annotées peuvent être utilisées (en partie) pour une comparaison systématique avec les décisions du système, et les paramètres peuvent être optimisés par des tests exhaustifs.

Ces méthodes sont basées sur la recherche de liaisons entre les variables descriptives et la variable ciblée, et utilisent diverses mesures statistiques pour faire émerger celle-ci et les expliciter. On retrouve donc dans ce calcul les mécanismes vus au chapitre précédent, mais systématisés et affinés puisque les systèmes d'apprentissage sont pour la plupart conçus pour tirer le meilleur parti du pouvoir prédicteur des descripteurs. Ils sont également conçus pour rechercher la stabilité du modèle, en évitant autant que faire se peut de faire émerger des configurations trop anecdotiques qui n'auraient pas de pouvoir prédictif suffisant.

Les techniques de fouille de données utilisent entre autres ces modèles d'apprentissage, mais en se concentrant sur le modèle construit plus que sur son pouvoir prédictif (ce dernier n'étant vu que comme une estimation de la qualité du modèle). Par contre, elles privilégient naturellement, comme on l'a dit, les techniques qui permettent une interprétation du modèle.

7.1.1.2 Méthodes

Les méthodes d'apprentissage supervisé que j'ai utilisées dans le cadre de mes travaux sont de différents types.

Les premières sont parmi les plus anciennes, et construisent des modèles à base de règles *symboliques*. Ces règles ont l'avantage d'être directement interprétables (à condition qu'elles soient en petit nombre), puisqu'elles prennent la forme d'une formule logique classique, qui combine des tests concernant les traits descripteurs (en partie gauche) et une valeur pour la variable-cible (partie droite). Il s'agit donc de règles du type :

Si l'item a les caractéristiques X_i alors la variable-cible a pour valeur Y

Chaque méthode différente produit un ensemble de règles à partir des données d'apprentissage, et les articule suivant plusieurs méthodes : soit dans une liste ordonnée (la première règle pouvant s'appliquer dans l'ordre donnant la réponse du système, voir 7.2.1, page 182), soit dans un arbre (arbre de décision, voir 7.2.2, page 183). Si l'avantage majeur de ces méthodes est leur interprétabilité (et dans une moindre mesure leur rapidité de calcul), elles ne sont, par contre, pas compétitives dans un ensemble de situations, notamment celles, de plus

en plus courantes en TAL, pour lesquelles le nombre de descripteurs est très élevé. En effet, les règles (ou les arbres de décision) ne peuvent articuler qu'un nombre réduit de descripteurs, et doivent donc opérer une sélection drastique parmi ceux-ci. Friedman résume leur limitation comme suit :

> *Finally, trees fragment the data. [...] Thus, each prediction involves only a relatively small number of predictor variables. If the target function is influenced by only a small number of (potentially different) variables in different local regions of the predictor variable space, then trees can produce accurate results. But, if the target function depends on a substantial fraction of the predictors everywhere in the space, trees will have problems.*
>
> Friedman (2006)

On verra que cette sélection et cet isolement des descripteurs est en fait un avantage quand on cherche justement à identifier ceux qui sont les plus pertinents et qui ont le plus d'influence sur la variable visée. Mais elle rend cette méthode impropre à la gestion de descripteurs de bas niveaux (comme les unités lexicales ou les n-grammes de caractères d'un texte) qui sont, par nature, très nombreux et ne portent chacun qu'une infime partie de l'information nécessaire pour décider de la classe d'un item.

Par opposition à ces systèmes interprétables, j'ai également été amené à utiliser des méthodes *numériques* ou *probabilistes*. En lieu et place de l'isolement des variables des méthodes symboliques, ce type de système articule les descripteurs par le biais de pondérations qui traduisent leur influence dans le processus de décision. Par exemple, les systèmes bayésiens fonctionnent par l'estimation d'une probabilité conditionnelle qui relie numériquement chaque descripteur à chaque classe, et la décision finale se fait sur la base d'un regroupement de ces probabilités individuelles pour arriver à une estimation probabiliste de l'appartenance à une des classes visées. Ce sont les modèles de ce type qui sont les plus utilisés à l'heure actuelle, car leur puissance de prédiction a été démontrée dans la plupart des situations que rencontre le TAL moderne. Ils sont notamment tout à fait capables de traiter de grandes quantités de descripteurs. La méthode à laquelle nous avons le plus souvent fait appel pour ce type d'utilisation dans l'équipe (comme de nombreuses autres en TAL) est un classifieur fonctionnant sur la base du principe d'entropie maximale (Berger *et al.*, 1996).

Le choix de cette méthode revient en fait à Assaf Urieli, qui s'est plongé dans la mécanique mathématique de ce modèle pour l'utiliser à des fins d'étiquetage et d'analyse syntaxique (voir 7.3.4). Le principe général de cette technique est de proposer une estimation de la distribution de probabilités des descripteurs associés aux données. Ces descripteurs sont en fait vus comme des contraintes qui régissent l'attribution d'une classe (ici, une étiquette morphosyntaxique, ou une relation de dépendance syntaxique). En procédant par itérations, l'algorithme calcule donc les pondérations associées à chaque type de descripteur (ici, les catégories possibles pour chaque mot, celles de ses voisins, sa position dans la phrase, etc.), en cherchant à maximiser l'entropie globale de la distribution, en respectant les contraintes du corpus d'apprentissage.

C'est donc tout naturellement que nous avons privilégié cette méthode particulière pour d'autres applications (voir 8.3, page 210), alors que d'autres méthodes auraient été envisageables sans changer la philosophie globale ni la nature et le nombre des descripteurs.

La complexité mathématique de ces méthodes est une des causes du creusement du fossé entre la linguistique classique et le TAL moderne sur lequel je reviendrai au prochain chapitre. En plus de la difficulté à comprendre les principes fondamentaux de ces méthodes, le modèle

qu'elles produisent n'est pas intelligible comme le sont les règles symboliques puisqu'il s'agit, dans le meilleur des cas, d'un ensemble de scores de coefficients associés à chaque descripteur indiquant sa plus ou moins grande influence dans le choix d'une valeur particulière pour la cible. Il ne faut donc pas espérer de ces méthodes qu'elles produisent une connaissance exploitable autrement que dans leur application directe. Autrement dit, elles prédisent mais n'expliquent pas, pour reprendre la formule de René Thom (1993).

L'utilisation concrète de ces modèles peut généralement se faire assez simplement, le nombre de paramètres à régler étant assez limité. Par contre, une des opérations importantes à effectuer en amont est la sélection des descripteurs. Malgré leur capacité à ingérer de grandes quantité de traits, les performances globales de ces méthodes ne sont pas systématiquement meilleures quand on injecte de façon cumulative de l'information pour chaque item. Par exemple, certaines de ces méthodes (notamment les classifieurs bayésiens) sont prévues pour fonctionner avec des descripteurs indépendants (au sens statistique). Ainsi, si plusieurs descripteurs représentent peu ou prou la même information ou sont fortement corrélés, les décisions du système seront artificiellement influencées plus fortement par celle-ci.

Heureusement, un des avantages de la méthode par entropie maximale est de ne pas exiger une telle indépendance, et donc de pouvoir gérer efficacement la redondance entre les descripteurs. Mais en aucun cas on ne peut aveuglément additionner les traits en espérant systématiquement une amélioration de la prédiction. Il est donc nécessaire d'utiliser différentes mesures statistiques en amont pour sélectionner les traits qui ont le meilleur potentiel de discrimination, et qui seront fournis en entrée au système. Cette phase ajoute encore à la complexification et à l'éloignement d'une compréhension du processus véritablement à l'œuvre. Par contre, en tant que telle elle il s'agit d'un moyen de mesurer la pertinence d'une caractéristique linguistique des items à traiter, qui peut avoir un intérêt dans la compréhension d'un phénomène, comme le font les méthodes statistiques vues au chapitre précédent.

7.1.2 Apprentissage non supervisé

L'apprentissage supervisé nécessite de cibler une variable particulière qui sera la sortie du système en phase d'exploitation. Les modèles fournis ne concernent donc que la « découverte » de régularités concernant les autres descripteurs permettant de prédire cette unique variable.

Il existe d'autres méthodes d'apprentissage, dites non supervisées, qui recherchent des régularités et des configurations globales des données sur lesquelles elles sont appliquées. Deux types de méthodes sont les plus connues dans ce domaine : le *clustering* (ou partitionnement de données) et les règles d'association.

7.1.2.1 Clustering

A la différence de la classification, le *clustering*[1] ne suppose pas l'existence de catégories prédéfinies des individus, et ne se base donc que sur des variables descriptives. Ces méthodes cherchent à mettre en évidence une organisation des données en classes (distinctes ou hiérarchiques), sur la base d'une relation de proximité (ou à l'inverse de distance) entre les individus. Cette proximité est basée sur un calcul de partage des propriétés exprimées par les descripteurs. Quelle que soit la méthode, deux individus dont les descripteurs sont à peu près identiques seront considérés comme proches, et donc comme faisant partie de la même classe.

1. J'utilise ici le terme anglais, puisqu'en français les termes *classification* et *catégorisation* sont dans l'usage des synonymes.

Deux individus dont toutes les variables sont nettement distinctes seront a priori des membres de classes distinctes. Mais le résultat final prendra systématiquement en compte l'ensemble de la collection étudiée.

Ces techniques reposent donc entièrement sur cette définition d'une mesure de similarité entre les individus. Il existe des distances par défaut, comme la distance euclidienne ou la distance de Manhattan qui sont des moyens classiques de synthétiser en une seule valeur numérique l'ensemble des différences existant entre les valeurs des variables pour un couple d'individus. Il est bien entendu possible de définir une distance plus complexe, notamment en pondérant chaque variable.

Une fois cette distance établie, les algorithmes de clustering fonctionnent par passes successives jusqu'à proposer une répartition de l'ensemble des individus dans des classes. Par contre, ces classes n'ont généralement pas d'interprétation directe en termes de configurations de variables, il est ensuite nécessaire de rechercher des relations entre l'appartenance à une classe et celles-ci, ce qui nécessite généralement un nouvel appel à des méthodes d'investigation (statistique ou par apprentissage). Néanmoins, ces méthodes produisent des informations qui peuvent être pertinentes sans que l'on cherche immédiatement à interpréter les classes. On peut notamment avoir un aperçu général de l'homogénéité de la collection, en fonction des distances obtenues à l'intérieur de chacune des classes (sont-elles resserrées ou non), et entre elles (sont-elles proches ou nettement distinctes). Pour certaines méthodes, notamment la classification hiérarchique, il est également possible d'observer la répartition au niveau des individus, et de repérer ceux qui sont les plus atypiques (voir 7.2.3, page 184).

Ce type de méthode rejoint donc l'étude de la répartition de la population à l'aide d'une analyse factorielle, dont les axes principaux proposent également une répartition des individus. Par contre, le *clustering* propose une partition explicite, qui n'est pas sujette aux problèmes de lecture des cartes factorielle.

7.1.2.2 Règles d'association

Le principe des règles d'association est de repérer les cooccurrences régulières de valeurs pour certains attributs. Le formalisme est particulièrement adapté à l'analyse de listes (le prototype des applications de cette méthode est l'analyse pour le marketing de listes d'achats, pour identifier les produits régulièrement achetés ensemble par des clients). Les règles obtenues ont la forme générique suivante :

Si l'item a les caractéristiques X_i, alors il possède également les caractéristiques Y_j

Il s'agit d'une méthode non-supervisée et qui ne vise donc pas une variable particulière, bien qu'on puisse ne sélectionner que les seules règles impliquant une variable que l'on souhaite étudier plus précisément.

Les règles d'association ainsi obtenues sont directement interprétables, comme le sont les règles des systèmes d'apprentissage supervisé symboliques, mais elles peuvent, comme elles, apparaître en très grand nombre. Cette méthode est donc particulièrement adaptée en TAL à l'extraction d'indices cooccurrents pour une structure linguistique particulière (voir 7.2.4 et 7.3.1).

7.1.3 Prédiction structurée

Si la situation dominante dans le domaine de l'apprentissage automatique est celle de la prédiction de la valeur d'une variable unique (qu'elle soit qualitative ou quantitative), certains

problèmes du TAL se traduisent par le besoin de résultats plus complexes.

François Yvon le résume ainsi, expliquant notamment par ce problème le manque d'enthousiasme des linguistes face aux techniques par apprentissage :

> *Despite their success, classification approaches have been subject to criticism, especially regarding the over-simplistic models of language they often imply. [...] The poverty of output representation is easier to see : many linguistic analysis tasks require more complex output than classes, such as graphs, trees, or (recursive) attribute-value structures. [...] In such situations, classification models cannot be used in isolation, and need to be complemented with an explicit modelling of dependencies found in output structures.*
>
> Yvon (2002)

C'est le cas de l'étiquetage syntaxique par exemple, qui ne se contente pas d'attribuer une classe à chaque unité, mais un ensemble de relations entre celles-ci. D'autres cas similaires sont légion en TAL (identification des coréférences, étiquetages sémantiques divers, etc.). Une évolution récente des techniques d'apprentissage vise donc spécifiquement à traiter ce genre de problèmes, en développant des méthodes spécifiques afin notamment de tenir compte de la nature de la structure visée (Smith, 2011; Daumé, 2006). Ces différentes méthodes et problématiques se regroupent sous le terme d'apprentissage structuré [2].

Le phénomène structurel que l'on cherche à capter par ces méthodes est ainsi décomposé en informations élémentaires, chacune correspondant grosso modo à une variable, ramenant ainsi l'apprentissage à une situation classique (par exemple, prédire pour chaque paire de mots d'une phrase si ceux-ci sont reliés syntaxiquement ou non, puis pour chaque paire de mots reliés, quelle est le rôle syntaxique exprimé auquel correspond cette relation). Le problème pris globalement consiste alors à prendre en considération, généralement dans un second temps, l'ensemble de ces informations locales pour tenir compte des interrelations qu'elles entretiennent (donc de la structure syntaxique de la phrase entière).

Cette phase consiste à définir et injecter dans le système des contraintes spécifiques, en interdisant ou en pénalisant certaines prédictions du système. Ces contraintes peuvent être d'ordre purement structurel, et exprimer une organisation logique des informations élémentaires (interdire les relations symétriques et les cycles) mais également d'ordre linguistique et correspondre à des connaissances sur le phénomène visé (une relation sujet au plus par verbe).

Sur le plan positif, ces nouvelles approches permettent donc en théorie un nouveau niveau d'interaction entre des approches quantitatives lourdes et des connaissances linguistiques, généralement exprimables de façon simple sous la forme de règles logiques régissant les sorties du système. Cette possibilité pour la linguistique d'intégrer des connaissances formalisées est bien entendu plus satisfaisante que la construction d'une simple collection d'exemples pour le corpus d'apprentissage.

Malheureusement, sur un plan plus négatif, la complexité énorme de ces systèmes et la prédominance des questions relatives à sa mécanique propre (l'articulation des différents systèmes d'apprentissage local, le paramétrage de ceux-ci, et la définition des contraintes purement structurelles) accroissent à mon avis encore plus la distance entre les deux mondes. De plus, il semblerait que les efforts soient pour l'instant essentiellement déployés pour des

2. Je remercie Philippe Muller de l'IRIT pour m'avoir éclairé sur l'apprentissage structuré, et avoir su dégager ses grands principes sommairement présentés ici

applications directes et de bas niveau du TAL (segmentation, étiquetage, repérage d'entités nommées, etc.) peu attrayantes par rapport à d'autres phénomènes plus complexes.

Dans ce champ encore plus que dans d'autres, il semblerait de plus que la linguistique soit montrée surtout comme un domaine soulevant des problèmes techniquement intéressants, et les modèles et les connaissances sur le langage comme une source inépuisable de nouvelles structures complexes.

7.2 Fouille de données langagières annotées

Je vais reprendre dans les sections qui suivent quelques-unes des analyses présentées dans le chapitre 6, et montrer comment l'utilisation des techniques de fouille de données permet la mise au jour d'informations concernant des données langagières annotées.

L'intérêt de ces techniques est assez direct : l'application de ces méthodes est rendu très aisée par la disponibilité d'outils très conviviaux comme la plate-forme Weka (Witten et Frank, 2005), et les modes de présentation des modèles obtenus les rendent facilement interprétables.

7.2.1 Extraction de règles pour la caractérisation des types de structures énumératives

Le premier exemple concerne l'extraction de règles à partir des données du projet Annodis concernant les structures énumératives (SE). Les règles présentées ci-dessous ont été extraites par la méthode RIPPER (Cohen, 1995), qui identifie des règles conjonctives à partir des descripteurs. Il s'agit d'une méthode par apprentissage, capable d'être utilisée pour prédire une variable qualitative (même si cet objectif n'est pas celui visé ici) : la variable ainsi ciblée est le type des structures énumératives. Les variables descriptives utilisées sont celles qui sont présentées en section 6.2 (page 144), les règles obtenues sont présentées dans la table 7.1.

Chaque règle obtenue est présentée sous la forme classique : la partie gauche est une conjonction de comparaisons entre une variable descriptive et une valeur (un seuil numérique ou une catégorie nominale), et la partie droite est la valeur déduite pour le type de SE. Les indications numériques à droite de chaque règle donnent deux valeurs : la première est le nombre de cas correctement déduits par la règle, et la seconde le nombre d'erreurs commises par son application (le rapport autorisé entre les deux est un des paramètres à fixer lors de l'utilisation de cette méthode). Ces deux valeurs sont calculées a posteriori sur les données utilisées pour la phase d'apprentissage, ici l'ensemble des structures énumératives du corpus Annodis. Cette suite de règles est destinée à être utilisée de façon séquentielle, leur ordre est donc important puisque la première règle applicable détermine la réponse du modèle.

La première règle indique par exemple que si le nombre de paragraphes sur lesquels s'étend une SE est supérieur ou égal à 6, et si le nombre de mots dépasse 837, alors il s'agit probablement d'une SE de type 1 (90 cas positifs, 11 exceptions).

L'application de cette méthode repose sur plusieurs paramètres, qui correspondent essentiellement à la définition du taux d'erreur maximal et de la généralité de chacune des règles. Comme tous les systèmes symboliques d'apprentissage, il est nécessaire en effet que les règles correspondent à une généralisation à partir des données utilisées, et non pas à une description très précise mais spécifique à ces seules données (ce que l'on appelle l'*overfitting*, ou sur-spécialisation d'un modèle). La procédure d'induction contient donc une phase de simplification des règles visant à en réduire la taille et le nombre jusqu'à arriver à un résultat satisfaisant les contraintes fixées par les paramètres. Le nombre de règles obtenues est donc

```
1: (N_Par ≥ 6) and (N_mots ≥ 837) ⇒ Type=1 (90/11)
2: (N_Par ≥ 6) and (Indices ≤ 5) ⇒ Type=1 (17/4)
3: (N_Par ≥ 2) and (Corpus = WIKI) and (N_Items ≤ 3) and (N_mots ≥ 413) ⇒ Type=1 (11/4)
4: (N_Par ≥ 2) ⇒ Type=3 (124/21)
5: (N_Par ≥ 1) and (Amorce = Non) ⇒ Type=3 (32/1)
6: (N_Par ≥ 1) and (Indices ≤ 5) and (N_mots ≥ 157) and (Indices ≤ 4) ⇒ Type=3 (22/6)
7: (N_Par ≥ 1) ⇒ Type=2 (192/27)
8: (Corpus = WIKI) and (N_Items ≥ 4) and (Indices ≥ 7) ⇒ Type=2 (11/2)
9: (Corpus = WIKI) and (N_mots ≤ 40) ⇒ Type=2 (8/1)
10: ⇒ Type=4 (336/11)
```

TABLE 7.1 – Règles apprises par RIPPER pour le typage des SE

fixé par le déroulement du programme, et dépend à la fois des données et du paramétrage. Dans l'exemple présenté ici, nous avons utilisé les paramètres par défaut proposé par l'implémentation de RIPPER dans la plate-forme Weka.

Plusieurs informations peuvent donc être trivialement obtenues à la lecture de ces règles :

- le type le plus fréquent est le type 4 (SE intra-paragraphique), comme l'indique la dernière règle (10), qui s'applique par défaut si aucune des précédentes n'a été déclenchée ;
- les types sont en partie définis sur la base des variables qui ont été fournies en entrée du système, il est donc normal (et rassurant) que certaines règles retrouvent cette évidence, comme la quatrième qui indique que les SE de type 3 s'étalent sur plus d'un paragraphe ;
- les autres règles mettent au jour des caractéristiques des SE qu'on avait pu repérer par investigation précise avec des mesures statistiques :
  - les SE de type 1 sont les plus longues (règle 1) ;
  - les SE de types 1 et 2 sont plus fréquentes dans le corpus Wikipedia (règles 3, 8 et 9) ;
  - les SE de type 2 sont celles qui ont le plus d'items et d'indices (règle 8) ;
  - les SE de type 3 ont moins d'amorces (règle 5).

L'idée de cet exemple n'est bien entendu pas de construire un système de typage automatique des structures énumératives : les types qui ont été définis se basent sur des caractéristiques différentes de celles qui ont été étudiées ici, et de toute façon chaque type est clairement décrit et identifiable. Il s'agit donc d'un cas classique de fouille de données basé sur une méthode d'apprentissage supervisé afin de découvrir des liaisons entre caractéristiques, ici en focalisant la variable étudiée comme étant la réponse souhaitée du système.

7.2.2 Arbres de décision pour étudier l'impact du sexe du médecin sur les consultations

Une autre méthode d'apprentissage automatique très célèbre pour sa facilité d'interprétation est celle des arbres de décision (Quinlan, 1993). Comme la méthode précédente, le modèle calculé se présente sous la forme de règles, mais cette fois-ci organisées hiérarchiquement de façon plus complexe qu'une liste.

L'exemple présenté ici s'appuie sur les données du projet Intermede, visant la description des caractéristiques linguistiques de la consultation liées au sexe du médecin. Les descrip-

teurs utilisés sont listés en section 6.4.3 (page 160), mais nous en avons exclu les variables extralinguistiques (âge, sexe, catégorie sociale et type de consultation) pour nous concentrer uniquement sur les faits langagiers.

L'arbre qui en résulte (utilisant la méthode C4.5) est présenté en figure 7.1. Comme pour les règles de l'exemple précédent, le paramétrage de cette méthode consiste à fixer le taux d'erreur de chaque décision prise, ainsi que la couverture minimale (nombre de cas concernés).

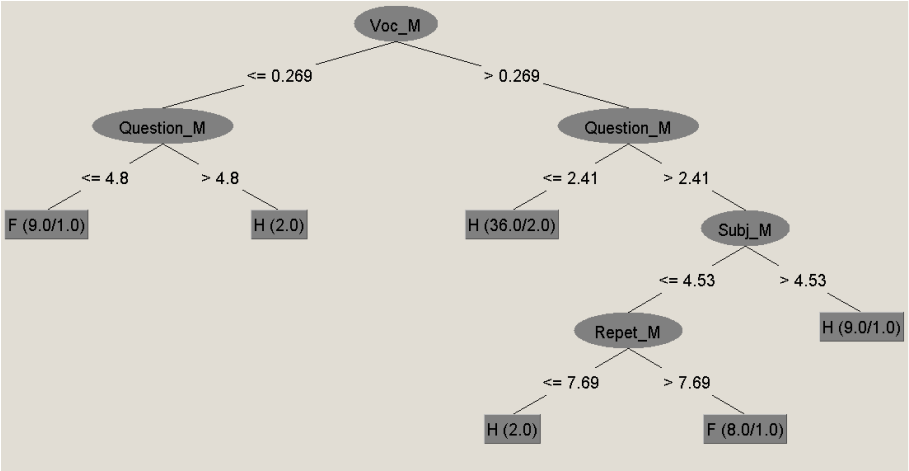

FIGURE 7.1 – Arbre de décision pour caractériser le discours du médecin en fonction de son sexe

Cet arbre indique la succession de tests à appliquer à une consultation pour en déduire le sexe du médecin : en partant de la racine, chaque nœud correspond à une dichotomie portant sur une variable (en fixant un seuil pour les variables numériques). Les feuilles correspondent à une décision pour la variable ciblée, indiquant comme précédemment le nombre de cas positifs et d'exceptions atteints par la règle.

On peut voir que la caractéristique la plus discriminante est la part de vocabulaire spécifique du médecin (*Voc_M*) : comme on l'avait vu ce sont les médecins hommes qui partagent le moins d'occurrences avec leur patient. De même, le nombre de questions posées par le médecin (*Question_M*) est un bon prédicteur (les médecins femmes en posent plus que les hommes), ainsi que l'expression de leur subjectivité (*Subj_M*) et leur propension à répéter les paroles du patient (*Repet_M*).

7.2.3 *Clustering* des appels de citation

Pour étudier les fonctions des citations bibliographiques dans les articles de SHS (projet Rhecitas, voir section 5.2.3, page 126), nous avions recherché, dans les contextes immédiats des appels de citation, un ensemble de caractéristiques, dont voici la liste :

- position relative dans le texte ;
- position de l'appel dans la phrase (début, milieu ou fin) ;
- présence ou non de parenthèses autour de l'appel ;
- présence de l'appel dans une séquence, ou de façon isolée ;
- en cas d'insertion de l'appel dans la phrase, fonction syntaxique de celui-ci (sujet, objet, dans un syntagme prépositionnel) ;

– cooccurrence avec des pronoms de première personne.

A la différence des deux situations précédentes, nous ne disposons pas à ce stade d'une typologie préétablie, et donc pas d'une variable-cible qu'un processus d'apprentissage classificatoire permettrait de viser. L'objectif ici est donc celui d'une observation de la répartition des citations en fonction de leurs caractéristiques. Si l'on peut mesurer individuellement la distribution des valeurs de chacune des variables descriptives par des méthodes statistiques, il est également possible d'utiliser une méthode d'apprentissage non supervisé pour faire émerger des régularités.

J'ai donc appliqué à ces données une méthode de *clustering*, fondée sur une distance entre les citations individuelles calculée sur la base du partage des caractéristiques ci-dessus. Plus précisément, j'ai utilisé ici la méthode *Expectation Maximization* implémentée dans Weka (Witten et Frank, 2005). Cette méthode a notamment un avantage sur d'autres, en ce sens qu'elle ne nécessite pas de fixer au préalable le nombre de classes entre lesquelles on souhaite répartir les différents individus de la collection de données.

Cette technique a permis de faire ressortir deux principales classes de citations :

– la première regroupe les appels de citation situés entre parenthèses, généralement en fin de phrase et souvent regroupés en énumération, positionnées essentiellement dans le premier tiers d'un article. Par exemple :

> *Plusieurs études ont montré que les mères favorisent les phrases courtes (Brown et Bellugi, 1964 ; Drach, 1969 ; Lord, 1975 ; Moerk, 1975 ; Nelson, 1973 ; Newport, 1975 ; Phillips, 1973 ; Sachs, Brown et Salerno, 1972 ; Shatz et Gelman, 1973 ; Snow, 1972, 1977).*

– la seconde concerne ceux qui sont insérés dans la phrase et qui sont le plus souvent isolés :

> *En cela, nous allons dans le sens des considérations de J.-P. Bronckart (1994) sur la double nature du genre.*

Cette répartition semble correspondre à ce que (Swales, 1990) qualifie de citations *integral versus non-integral*.

Cette première approximation, ainsi que la rareté des indices de plus haut niveau (pronoms, positions syntaxiques privilégiées, etc.) nous a conduits à la définition d'une typologie directement basée sur ces constatations, et l'hypothèse (en partie validée) que les citations les moins bien intégrées (première classe) correspondent à des citations d'arrière-plan. Voir Tanguy *et al.* (2009) pour plus de détails.

7.2.4 Règles d'association sur les indices des structures énumératives

Je vais revenir une fois de plus sur les données du projet Annodis, pour analyser plus finement cette fois les différents types d'indices qui signalent les structures énumératives, considérant que nous n'avons pas pu nous contenter des simplifications qu'exigeait l'examen global des caractéristiques de celles-ci (comme vu au chapitre précédent). La recherche de configurations de tels indices est un contexte idéal pour appliquer les règles d'association : à chaque SE est en effet associée une liste de longueur variable d'indices typés. La configuration globale est celle que l'on a vue sous la forme de treillis en section 5.1.2 (page 114) ; nous allons voir ici comment en extraire des informations plus synthétiques concernant les types d'indices.

Les règles d'association permettent en effet de retrouver les configurations les plus récurrentes, en faisant émerger les associations de type d'indices les plus régulières. Elles ressemblent de fait aux règles de décisions vues en section 7.2.1, mais ne visent pas dans leur conclusion une variable précise, et reposent sur la simple présence des indicateurs (ici un indice de SE) et pas sur des valeurs.

Les règles suivantes ont ainsi pu être extraites des données Annodis à l'aide de l'algorithme *Apriori*. Les règles sont présentées dans la table 7.2 par ordre décroissant de confiance, avec un seuil minimal de 75% (la confiance d'une règle est sa précision, donc le pourcentage des cas correspondant à la prémisse pour lesquels la conclusion est valide).

1. `Ponct. d'item ⇒ Ponct. d'amorce`
2. `Ponct. d'amorce ⇒ Ponct. d'item`
3. `Ponct. d'item ET Ponct. d'amorce ⇒ Indice lexical d'amorce`
4. `Titre ⇒ Indice lexical d'amorce`
5. `Indice lexical de clôture ⇒ Indice lexical d'amorce`

TABLE 7.2 – Règles d'association pour les types d'indices des structures énumératives

Les règles symétriques 1 et 2 indiquent que lorsqu'un indice ponctuationnel se trouve dans l'amorce (généralement un « : », on en trouve également dans les items (puces et/ou points-virgules), et vice-versa. Elles reflètent donc la cohérence des schémas de marquage ponctuationnel des SE, comme ceux visibles dans les exemples 2 et 4 de la table 5.1 (page 116).

Les règles 3 et 4 indiquent que les SE de type 1 (*i.e.* celles dont les items sont des sections titrées, règle 4) et celles qui sont marquées par des schémas ponctuationnels (essentiellement celles de type 2, règle 3) ont des indices lexicaux dans leur amorce. Il s'agit en fait d'une conséquence directe de faits déjà connus, à savoir que les SE de type 1 et 2 sont très souvent amorcées, et que les amorces ont pratiquement systématiquement un indice lexical (c'est le cas des exemples 1 et 2 de la table 5.1).

La règle 5 relie l'existence d'un indice de clôture à celle d'un indice lexical dans l'amorce. Mais, comme les amorces et les clôtures sont quasi-systématiquement dotées de tels indices, cette règle traduit surtout la présence systématique d'une amorce dans les SE dotées d'une clôture, et qu'il n'y a donc pas d'alternance véritable entre ces deux éléments.

On voit donc que ces règles ne font apparaître que des informations déjà obtenues par des observations statistiques classiques des SE du corpus, sans les atteindre toutes. De plus, elles ne sont pas capables de repérer les configurations maximales stables de traits observées à la lecture des treillis de la section 5.1.2 (page 114). Néanmoins, il faut reconnaître à ces règles la grande facilité de leur obtention qui en fait, comme pour les cas précédents, une bonne première approche d'un ensemble de données à observer.

7.3 Applications à base d'apprentissage automatique

Les travaux que je présente dans cette seconde partie font un usage plus intensif et plus central des méthodes quantitatives en TAL, et plus précisément de l'apprentissage automatique. On verra donc qu'à chaque fois les usages correspondent à un objectif plus précis que l'investigation d'un phénomène linguistique, même si cette phase est bien entendu également présente à chaque fois.

Comme on le verra, ici encore les situations sont très variées, tant par les données et les phénomènes visés que par les descripteurs sur lesquels se basent les procédures d'apprentis-

sage. Le point commun aux travaux qui sont détaillés ici est, en plus du caractère relativement récent de leur apparition dans mon travail de recherche, la volonté d'utiliser ces techniques en envisageant dès le début un retour vers le traitement linguistique, et l'articulation avec des questions plus centralement dirigées vers les données et les descriptions linguistiques. Ces questions peuvent, là encore, concerner la description d'un phénomène, ou plus largement éclairer le rôle des descripteurs linguistiques dans les processus. De ce fait, j'insisterai bien moins ici sur le fonctionnement des méthodes elles-mêmes que sur leur utilisation.

7.3.1 Repérage des segments d'obsolescence (thèse de Marion Laignelet)

Le travail que je présente ici est celui de la thèse de Marion Laignelet (Laignelet, 2009), dirigée par Marie-Paule Péry-Woodley et dont j'ai assuré la deuxième partie de l'encadrement (après le départ de Didier Bourigault). Cette thèse s'est déroulée dans le cadre d'une convention CIFRE avec la société d'édition Initiales, dont une des activités principales concernait des fiches encyclopédiques. L'objectif appliqué de cette thèse était donc l'assistance à la mise à jour de ces fiches, et plus précisément le repérage automatique des passages textuels susceptibles de nécessiter une telle mise à jour (les *segments d'obsolescence*). De façon prototypique, de tels segments sont ceux qui incluent des informations factuelles datées ou dont l'évolution est aisément prévisible (telles que les informations géopolitiques ou économiques dans la fiche décrivant un pays, par exemple « *Nicolas Sarkozy, président de la république française* »), mais ils peuvent également prendre des formes plus variables, et n'être identifiables que par des formulations liées à l'aspect ou à la modalité par exemple.

L'identification de ce type de segment est donc une tâche difficile, et leur repérage ne peut être abordé qu'à travers des indices linguistiques de différents types, fonctionnant de façon conjointe. Le travail de Marion Laignelet a consisté, à travers une étude approfondie sur un corpus annoté, à proposer dans un premier temps une collection d'indices linguistiques a priori pertinents pour la notion d'obsolescence, et à les projeter automatiquement sur les phrases du corpus. Ces indices, au nombre de 146, correspondent à différents niveaux d'analyse (lexique et syntaxe, mais aussi discours, avec la prise en compte de la position relative des phrases et leurs rapports avec les titres de sections) ont nécessité le développement d'outils et de ressources spécifiques (lexiques spécifiques et patrons morphosyntaxiques, etc.) implémentés dans la plate-forme LinguaStream (Bilhaut et Widlöcher, 2006).

L'étape d'analyse a consisté ensuite à mesurer la pertinence de ces indices, notamment à travers la mesure de corrélations avec le marquage manuel de l'obsolescence de chacune des phrases, mais aussi avec l'exploration de leur comportement en cooccurrence à l'aide de méthodes multivariées. Ces études ont permis de confirmer la pertinence de la plupart des indices, ainsi que la nécessité de les utiliser en faisceaux pour le repérage automatique des segments d'obsolescence. Ce dernier aspect, qui concerne donc le développement d'un classifieur, a fait appel à la technique des règles d'association (dans le cadre d'une collaboration avec François Rioult du laboratoire GREYC à Caen). Bien qu'initialement développées pour la fouille de données, ces règles peuvent également être utilisées pour des opérations de classification, suivant différentes méthodes similaires à celles des règles classiques en apprentissage automatique, comme le proposent Li *et al.* (2001). L'ensemble de la procédure et les résultats très encourageants obtenus sont présentés dans (Laignelet et Rioult, 2010; Laignelet *et al.*, 2010). Si la méthode par règles d'association a l'avantage, comme on l'a vu, de produire des résultats interprétables, l'inconvénient majeur est le grand nombre de règles ainsi produites, chacune ayant au final une faible couverture. Toutefois, les règles les plus productives per-

mettent de mettre au jour des combinaisons d'indices intéressantes, mélangeant notamment des indices de différents niveaux. C'est le cas par exemple de schémas récurrents dans lesquels le segment obsolescent est caractérisé par sa position en initiale de paragraphe, dans la portée d'un titre contenant une unité lexicale catégorisée, et contenant une valeur numérique, comme dans l'exemple ci-dessous (extrait du *Grand Universel Larousse*, entrée *Botswana*) :

> **x. Population**
> En raison de l'omniprésence à l'ouest du désert du Kalahari, la population se concentre pour l'essentiel dans l'est, le long du grand axe routier et ferré entre l'Afrique du Sud et le Zimbabwe et qui relie les principales villes du pays dont la capitale Gaborone (**195 000 hab.**).

Ce travail s'inscrit dans un type d'études empiriques sur un phénomène linguistique précis, visant à la fois sa caractérisation générale à partir de corpus annotés par l'identification d'indices linguistiques permettant son repérage, et le développement de méthodes automatiques à cette fin. La complexité des phénomènes, notamment ceux relevant de l'organisation du discours, nécessite l'identification de faisceaux d'indices, et donc la découverte de leurs combinaisons (voir par exemple Bouffier (2009) pour le cas des recommandations dans les textes médicaux). Je reviens sur cette méthodologie en section 8.4.2.

7.3.2 Traitement des rapports d'incidents aériens (collaboration avec la société CFH et thèse de Nikola Tulechki)

Le travail que je présente ici s'inscrit dans une collaboration de longue date avec la société CFH (Conseils en Facteurs Humains), dont la dernière concrétisation est la thèse de Nikola Tulechki (démarrée en janvier 2011) que j'encadre avec Marie-Paule Péry-Woodley et Éric Hermann (directeur de CFH) dans le cadre d'une convention CIFRE.

Une activité importante de CFH dans le cadre de la sécurité des transports aériens concerne l'étude et la classification de rapports d'incidents, une forme de retour d'expérience (REX) produite par les pilotes des compagnies aériennes ou les experts des instances de régulation, et qui décrivent en texte libre le déroulement d'un événement potentiellement dangereux survenu lors d'un vol. Ces rapports nécessitent, dans le cadre des politiques d'amélioration de la sécurité du transport aérien, une analyse postérieure par un expert et une classification en fonction d'une nomenclature précise, visant à faire ressortir explicitement les causes de l'incident, et donc à terme les façons de les prévenir. La quantité de rapports ainsi produits, et le manque de disponibilité des experts capables de les interpréter créent un besoin croissant de méthodes automatiques pour assister leur traitement, que les responsables de CFH ont su faire comprendre aux différentes instances de décision (compagnies aériennes et organismes de régulation du trafic aérien). En plus de ces objectifs imposés par la réglementation en vigueur, il apparaît clairement à l'ensemble des acteurs que ces données (structurées ou non) sont une source vitale de connaissances permettant d'anticiper des problèmes, et qu'il y a également un besoin dans leur analyse plus large.

7.3.2.1 Classification automatique des rapports : vers un système de prédiction structurée

Le premier travail (auquel je n'ai pas participé, ce sont mes collègues Didier Bourigault et Cécile Fabre qui l'ont initié) a consisté à développer un système de classification automatique

sur la base d'une extraction terminologique, afin d'identifier quels candidats-termes sont associés aux catégories de la nomenclature en vigueur, notamment la taxonomie ECCAIRS [3] (Hermann *et al.*, 2008). Le dispositif actuel est basé sur une méthode bayésienne qui estime, sur la base de rapports déjà classés, la probabilité qu'un terme soit un indicateur d'une catégorie, et qui utilise un système de seuils pour prédire les catégories les plus plausibles pour un rapport donné. Ce système permet à la fois de proposer des suggestions à l'expert qui classe les rapports, mais aussi de repérer des incohérences dans les bases de données existantes. Nous sommes actuellement en train d'expérimenter le remplacement de cette méthode par un classifieur par entropie maximale (grâce à la compétence récemment acquise dans l'équipe et au développement par Assaf Urieli d'un outil finalisé [4]). Cette évolution de la méthode statistique ne remet pas en cause le grand nombre de prétraitements spécifiques développés par les linguistes de CFH, qui projettent sur les textes des connaissances linguistiques vitales pour l'exploitation efficace de ces textes particuliers (truffés d'acronymes, de nombreux codes spécifiques, de termes anglais, etc.), allant de simples règles de réécriture à des ontologies partielles du domaine. En effet, les deux méthodes de classification utilisent en entrée des candidats-termes pertinents pour le domaine et normalisés. D'ailleurs, des expériences préliminaires viennent de confirmer le gain significatif de ce travail minutieux de prétraitement des données dans la tâche de classification. La valeur ajoutée de la méthode par entropie maximale est, elle, en cours de calcul.

Cette utilisation de méthodes par apprentissage entre en fait dans le cadre des problèmes nécessitant une prédiction structurée (voir 7.1.3). En effet, la complexité des nomenclatures visées par l'analyse de ces rapports se traduit par une classification en champs multiples (phase de vol, type d'événement, cause probable, dispositif de sécurité impliqué, etc.), ces champs entretenant entre eux des relations de dépendance. Bien que pour l'instant ces dépendances soient traitées par un ensemble de règles définies explicitement par les experts et appliquées en sortie de la classification automatique, nous envisageons de les intégrer plus au cœur du système lui-même en utilisant les techniques de la prédiction structurée.

Un dernier point en cours d'étude, et qui est rendu possible par le classifieur par entropie maximale, est l'intégration des méta-données dans le processus. Malgré le fait que ces rapports sont essentiellement composés d'un texte libre, ils sont généralement accompagnés d'indications génériques concernant des informations factuelles structurées comme le lieu, la date, le type d'appareil, les conditions météorologiques, etc. S'il est difficile, dans une approche probabiliste concentrée sur le seul contenu textuel, d'intégrer ces informations (puisqu'il faut alors en estimer la pondération relative), cela devient très simple en utilisant l'entropie maximale qui prend en charge la combinaison de descripteurs hétérogènes.

7.3.2.2 Calcul de la similarité entre rapports : un problème à plusieurs dimensions

Si la classification de ces rapports répond à une exigence facilement compréhensible pour la capitalisation et l'utilisation des connaissances, elle n'en n'est pas moins une vision réductrice de la complexité des données analysées. Les travaux de doctorat de Nikola Tulechki abordent précisément d'autres dimensions et envisagent des exploitations de ces documents par des approches moins normalisées. L'objectif applicatif de cette thèse est de permettre aux experts d'aborder ces collections avec d'autres outils qu'un moteur de recherche basé sur des champs

3. Voir `http://eccairsportal.jrc.ec.europa.eu/`
4. Baptisé CSVLearner, il est disponible à l'adresse suivante : `https://github.com/urieli/csvLearner`.

structurés, voire de permettre la mise au jour de facteurs de risques qui passeraient inaperçus sous ce seul prisme. C'est notamment ce genre d'information cachée que recouvre le terme de « signal faible », et qui constitue le graal des applications en fouille de données dans nombre de domaines appliqués.

Les travaux de Nikola Tulechki sont donc orientés vers l'analyse transversale des rapports d'incidents, dans plusieurs directions complémentaires. Si certaines se concentrent sur les modes d'expression des rapports (et notamment des marques de la subjectivité des rédacteurs, parfois très riches), d'autres se penchent sur des mesures d'une similarité plus globale entre les documents. Cette similarité peut bien entendu être définie sur la base des descriptions structurées, mais aussi concerner le contenu textuel, comme le font classiquement les modèles vectoriels des systèmes de recherche d'information.

Un premier usage direct concerne l'identification de rapports atypiques (*outliers*) que l'on peut aisément identifier par des méthodes de clustering ou de classification hiérarchique sur la base d'une similarité lexicale (Tulechki, 2011). Mais le point le plus intéressant de la réflexion concerne ici le croisement entre les différentes dimensions qui permettent de rapprocher les documents.

Le cas le plus marquant, et apparemment le plus efficace actuellement, est celui de la corrélation entre la similarité textuelle et l'ordre chronologique des incidents. Une des définitions des signaux faibles concerne en effet la notion de croissance rapide d'une source de problème qui peut être repérée (malheureusement généralement a posteriori) à travers des schémas d'occurrence temporelle. C'est ce qu'a proposé Nikola Tulechki sur la base documentaire publique de l'Aviation Safety Network, en proposant un mode d'accès et de visualisation qui positionne sur un graphique temporel les rapports similaires à un rapport-pivot choisi par l'utilisateur. L'application TimePlot qu'il a développée à cet effet est présentée en figure 7.2. On y voit précisément comment sont facilement identifiables dans la courbe du bas les différents rapports émis dans la période de l'éruption volcanique islandaise du printemps 2010. D'autres configurations pertinentes de répartition temporelle peuvent apparaître, notamment pour des phénomènes périodiques (conditions météorologiques ou autres). Cette application, basée sur un simple calcul de similarité lexicale, remporte un grand succès auprès des experts qui utilisent ces bases de données. On voit donc une fois de plus l'importance (ici pour la diffusion applicative d'une méthode) de la visualisation des données, notamment si elle est accompagnée comme ici d'une interface interactive permettant l'accès direct aux données. Le problème est tout autre si l'on souhaite sur cette base identifier automatiquement des configurations particulières.

Dans le même ordre d'idée, il est possible d'ajouter une dimension temporelle aux techniques de *clustering*, notamment en calculant l'évolution des classes induites au fil du temps. Les bases de données de rapports sont en effet en croissance continue, et l'observation de la modification des classes et des rapports atypiques est une façon opératoire d'aborder la caractéristique temporelle du signal faible.

L'hypothèse de travail que nous avons définie pour la suite de ce doctorat concerne précisément l'articulation entre la similarité textuelle et les autres dimensions descriptives de ces rapports. Si la corrélation entre le contenu textuel et les catégories de la nomenclature est postulée (et en grande partie vérifiée) par le processus de classification automatique présenté dans la section précédente, plusieurs questions restent en suspens. La principale concerne la mise au jour de similarités *orthogonales* à cette classification. Il s'agira donc dans un premier temps de faire la part des choses entre les éléments de textes (termes et autres indices lexico-syntaxiques) et les catégories, afin de neutraliser celles-ci et de calculer une similarité d'un

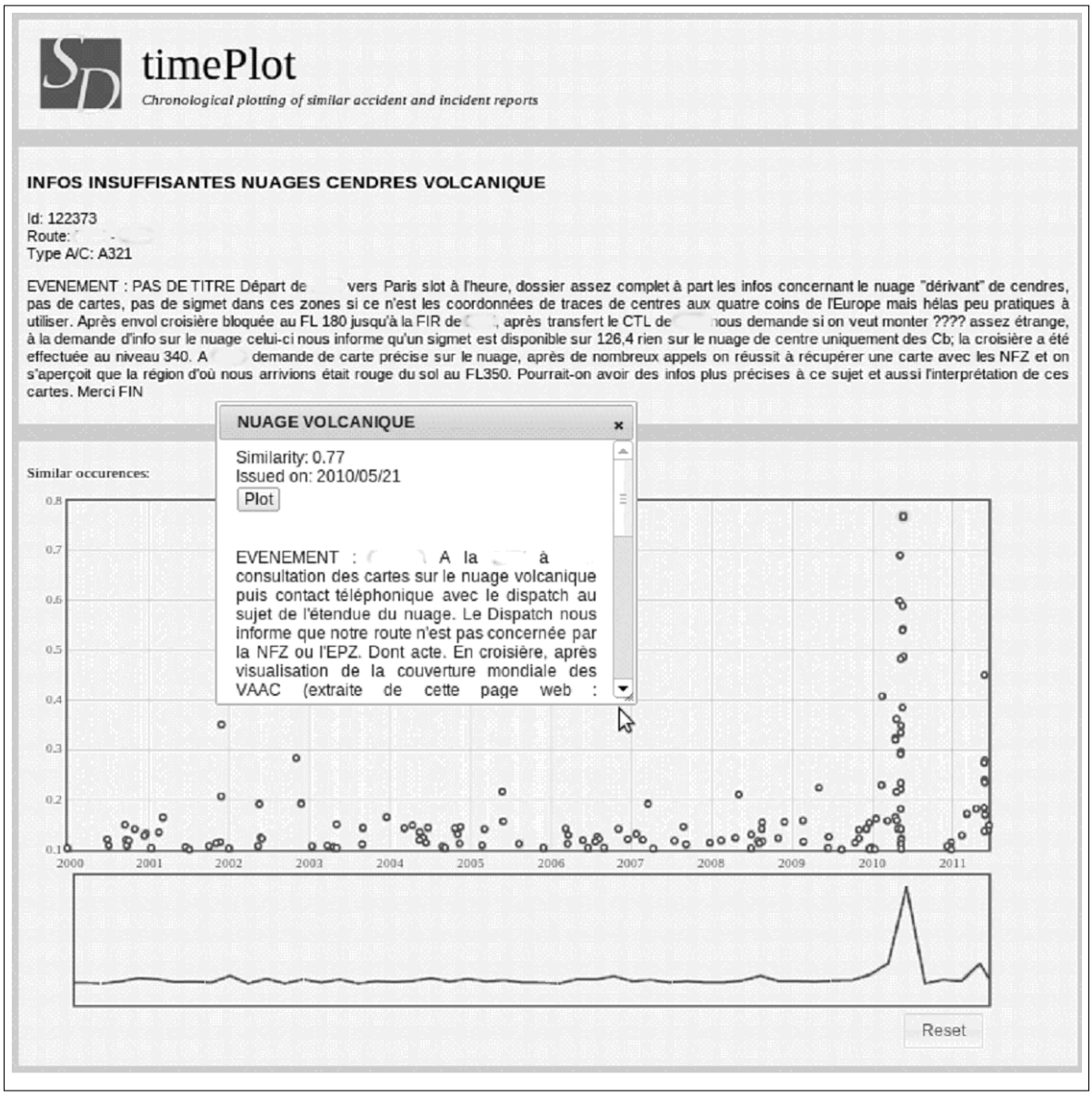

FIGURE 7.2 – Similarité des rapports d'incidents sur un axe chronologique
Le rapport-pivot sélectionné par l'expert est visible en haut de l'écran, et le reste de la collection est présenté sur le graphe temporel en bas (un clic sur un point de ce graphique permet de visualiser le contenu du rapport correspondant).

autre type sur la base de ce même contenu. Autrement dit, il s'agirait, comme le fait une analyse en composantes principales, d'identifier clairement la première dimension pour pouvoir examiner les autres. Sur ce même principe, nous allons chercher à identifier les éléments textuels générateurs de similarités entre les documents qui sont associables à d'autres données externes disponibles (lieu, temps, appareil, etc.). Le but final étant alors de faire émerger des dimensions pertinentes pour l'exploitation de ces données, passées jusqu'ici inaperçues à cause

de la prédominance d'un point de vue privilégié.

On le voit, ce travail fait donc un usage intensif de techniques quantitatives variées, afin de pouvoir exploiter au mieux la grande richesse des informations disponibles dans ces données.

7.3.3 Classification des requêtes en recherche d'information (projet CAAS et thèse de Simon Leva)

Le travail évoqué ici est celui de la thèse de Simon Leva (démarrée en octobre 2011) que je co-encadre avec Josiane Mothe de l'IRIT, et qui s'inscrit dans deux projets, CAAS et CITRI. Il s'agit de l'épisode le plus récent dans la série de travaux que j'ai menés en commun avec J. Mothe dans le domaine de la recherche d'information, et qui concerne à nouveau la caractérisation linguistique des requêtes, dans la lignée du projet ARIEL (voir section 6.4.2, page 159). Dans le cadre de ce travail, et à la différence des précédents qui se basaient sur l'environnement artificiel des campagnes TREC, nous avons la chance de travailler sur des données particulièrement intéressantes, à savoir les requêtes soumises par les utilisateurs au moteur de recherche du portail `revues.org` du CLEO (Centre pour l'Édition Électronique Ouverte, dirigé par Marin Dacos, voir `http://cleo.cnrs.fr`).

Ces données présentent de nombreux avantages :

- il s'agit de requêtes adressées à un moteur de recherche gérant une collection spécifique de documents (des publications en SHS), et donc écrites par des utilisateurs relativement homogènes (en majorité des universitaires et des étudiants), surtout lorsque l'on compare celles-ci à celles des moteurs de recherche généralistes sur le Web ;
- les documents nous sont en partie familiers, et même si de nombreuses publications concernent des disciplines très différentes de la nôtre comme l'histoire et l'anthropologie, elles nous sont tout de même plus proches que les documents techniques, comme les rapports d'incidents aériens. De plus, nous avons déjà eu l'occasion de travailler sur une partie de ces documents lors du projet RHECITAS ;
- la collaboration directe avec le CLEO nous permet d'avoir accès à des ressources dédiées très utiles : tout d'abord une identification des utilisateurs via leur adresse IP anonymisée, permettant notamment de considérer des sessions de requêtes, et d'étudier leurs reformulations par les utilisateurs, mais aussi des ressources liées aux documents, comme la structuration globale de la base documentaire (titres, auteurs, noms des revues, etc.) ;
- la possibilité à terme d'expérimenter et de déployer de nouvelles fonctionnalités sur le moteur afin d'en étudier l'impact sur les utilisateurs.

L'objectif des travaux menés dans ce cadre est la caractérisation automatique des requêtes des utilisateurs, en vue notamment d'améliorer la réponse du système de recherche d'information par un processus adaptatif. La première étape concerne l'établissement d'une typologie de ces requêtes, et l'identification de descripteurs permettant de calculer la catégorie d'une requête particulière. Les étapes suivantes concerneront le paramétrage adaptatif du moteur, et le profilage de documents permettant, le cas échéant, de sélectionner ceux-ci en fonction de leurs caractéristiques textuelles.

Les premiers travaux menés sur ces données nouvellement acquises nous ont permis de dégager des comportements différents dans la conception d'une requête adressée au portail `revues.org`. Sur la base d'une première observation d'un échantillon de quelques milliers de requêtes, on peut en effet définir des catégories non thématiques. Citons entre autres :

- des noms d'auteurs d'articles ("*Jean-Jacques Dufaure*") ;
- des titres d'articles ("*De l'illusion totémique à la fiction sociale*") ou de revues ("*dossiers*

d'archéologie") ;
- des thèmes génériques ("*esclavage*", "*bonheur au travail*") ou plus spécifiques ("*Recensement enquête démographique Afrique*"), parfois élaborés en séquence de reformulations successives ;
- des références complètes ("*Les espaces publics à Beyrouth in Geocarrefour vol 77.3*")
- des couplages thématique/auteur ("*formes élémentaires vie religieuses durkheim*", "*chiasme Merleau-Ponty*") ;
- des extraits de documents ("*2010, l'année où les catastrophes naturelles ont fait le plus de morts*").

On peut voir à travers ces exemples que les résultats attendus par les auteurs de ces requêtes vont varier d'un type à l'autre, allant de la recherche d'une publication précise à la nécessité de prendre en compte plusieurs documents pour établir un état de l'art ou cerner un nouveau champ d'étude. C'est notamment par ce biais que ce travail va tenter de rejoindre celui initié dans le cadre du projet RHECITAS sur le typage des documents par le biais de leurs références, présenté en 5.2.3 (page 126) ; ce sont ces aspects que visent plus précisément le nouveau projet CITRI.

Le repérage de ces différents types (une fois validés) passera par l'identification d'un certain nombre de traits calculables automatiquement à partir du texte des requêtes elles-mêmes. Même si ces données très particulières ne se prêtent pas à des analyses complexes, elles n'en ont pas moins un ensemble de caractéristiques dont la pertinence devra bien entendu être vérifiée.

Une première expérience en ce sens (le mémoire de master de Simon Leva) a concerné l'identification des variations entre deux types objectifs de requêtes, à savoir celles adressées à l'ensemble du portail `revues.org` et celles qui visent la collection d'articles ne correspondant qu'à une seule revue (le moteur qui les traite est en fait le même, mais du point de vue de l'utilisateur il s'agit bien de deux contextes différents). Simon Leva a ainsi pu identifier que les requêtes visant une revue particulière ont (significativement) tendance à être plus courtes et à utiliser un style plus compact (absence de déterminants, de signes de ponctuation et de prépositions) que celles adressées à l'ensemble du catalogue. Cette première expérience a donc surtout permis de valider l'influence du contexte (ici le point d'entrée dans le moteur de recherche) sur leur formulation. D'autres types de caractéristiques sont actuellement à l'étude dans une approche similaire (séquences et type de reformulations, utilisation de certaines fonctionnalités liées à l'interface du moteur, à terme nombre et types de documents consultés à la suite de la requête).

Les étapes suivantes de ce travail vont a priori chacune faire appel à des méthodes quantitatives. Un apprentissage non supervisé sur la base de caractéristiques linguistiques projetées sur les requêtes permettra de faire émerger des classes de requêtes qu'il faudra par la suite mettre en regard de considérations externes. Par la suite, une phase de typologie manuelle et un apprentissage supervisé permettra de proposer un prototype de classification, qu'il faudra bien sûr corréler avec les autres composants du système (types de documents, moteur d'appariement et critères de classement des résultats).

7.3.4 Étiquetage et analyse syntaxique (thèse d'Assaf Urieli)

Le doctorat d'Assaf Urieli démarré en 2009 (que je co-encadre avec Marie-Paule Péry-Woodley) vise le développement d'un analyseur syntaxique statistique configurable, en envisageant son adaptation à différents besoins en TAL. Initialement motivée par le vide qu'a

laissé dans l'équipe la fin de la disponibilité de l'analyseur Syntex, cette thèse est également l'occasion de réfléchir à l'articulation d'un analyseur syntaxique avec les ressources injectables et les utilisations très variées qui en sont faites dans nos activités.

Techniquement, Assaf Urieli est actuellement en phase de finalisation de l'analyseur Talismane, composé en amont d'un module d'étiquetage morphosyntaxique et en aval d'un analyseur syntaxique en dépendances, tous deux entraînés sur le corpus annoté du French Treebank (Abeillé *et al.*, 2003). Les deux modules se basent sur un classifieur par entropie maximale (MaxEnt).

Le recours à un système par apprentissage est désormais incontournable au simple regard des coûts de développement d'un analyseur à base de règles, il ne doit pour autant pas éjecter les préoccupations plus linguistiques en dehors de la phase de conception et de paramétrage. Ainsi, les travaux d'Assaf Urieli vont permettre d'aborder un ensemble de questionnements que je vais évoquer ici.

Les premiers concernent l'intégration des ressources linguistiques dans l'analyseur. Si le minimum d'informations nécessaires pour effectuer une analyse syntaxique consiste en un lexique morphologique et une collection de phrases déjà annotées, on peut considérer l'injection dans la boucle de ressources d'un autre type :

- *Des ressources génériques*, issues de travaux descriptifs comme les tables du lexique-grammaire et les informations de sous-catégorisation qu'elles contiennent. Leur utilité pour l'analyse syntaxique a d'ailleurs été étudiée en détails dans le cas de Syntex par Frérot *et al.* (2003). C'est également le cas de lexiques plus spécialisés comme Verbaction (voir section 4.4.3.4, page 99), dont les couples morphologiques peuvent permettre d'infléchir les décisions de rattachement en considérant l'héritage de la structure actancielle d'un verbe vers celle du nom d'action qui lui est morphologiquement apparenté. Henestroza Anguiano et Candito (2011) envisagent également de telles injections dans leurs travaux en analyse syntaxique statistique, sous forme de corrections en sortie de l'analyse.
- *Des ressources spécifiques au corpus ou au domaine visé.* On a vu notamment dans le cas des rapports d'incidents aériens que la prise en compte des spécificités de surface d'un corpus est bénéfique, surtout lorsque celui-ci s'éloigne des caractéristiques du corpus d'entraînement. Rappelons que le French Treebank est un corpus exclusivement journalistique, dont les structures phrastiques sont très différentes de celles d'un corpus technique par exemple. Les connaissances spécifiques peuvent prendre plusieurs formes, allant de règles de segmentation particulières à des structures terminologiques préétablies, en passant par des classes lexicales et un pré-étiquetage en amont. Bien souvent dans le cas des corpus spécialisés, de telles ressources sont disponibles et, même si elles sont partielles, elles permettent aisément d'envisager une adaptation de la chaîne d'analyse.

Ces ajouts de connaissances peuvent se faire soit en amont dans les phases de segmentation ou d'étiquetage, soit en aval par des règles de réécriture des sorties, soit encore directement dans le processus d'apprentissage sous forme de prédicteurs locaux. Dans tous les cas, il est bien entendu souhaitable, voire nécessaire, d'avoir la maîtrise du processus complet, c'est pourquoi rien ne peut remplacer le contrôle complet sur l'outil, ce que ne permet pas par exemple l'utilisation d'une chaîne générique finalisée dont on ne maîtriserait que les seules données d'entraînement. Sur ce dernier point, deux pistes nous semblent intéressantes.

La première concerne l'intervention directe d'un expert linguiste dans l'ajout de corrections manuelles au sein du modèle, permettant une correction spécifique de certains cas. C'est

une méthode similaire à ce que faisaient Bouillon *et al.* (2000) pour l'étiquetage morphosyntaxique, en apportant des corrections manuelles au modèle markovien induit des données d'apprentissage, afin de corriger les principales erreurs identifiées. Certes, l'intervention directe dans la matrice de poids d'un modèle à entropie maximale n'est pas aussi aisée que pour un modèle markovien, dans lequel les probabilités de transition restent à peu près compréhensibles, mais une intervention correctrice du même type en sortie du calcul des probabilités d'attribution d'une classe reste envisageable.

La seconde consiste à ajouter une intervention lors de la phase d'entraînement du modèle, en suivant le paradigme de l'apprentissage actif (*active learning*), voir la présentation synthétique qu'en a faite Settles (2009). Le principe général de cette variation sur le thème de l'apprentissage automatique supervisé est d'impliquer un expert du domaine lors de la phase de construction du modèle, en se basant sur un corpus non annoté. Par itération, le système est appliqué à des données non annotées, et certains de ces items sont sélectionnés en fonction de la mesure de confiance du modèle. Ces items « difficiles » sont présentés à un expert du domaine qui les catégorise manuellement, et sont ajoutés ensuite au corpus d'entraînement. Ce principe a déjà été utilisé pour plusieurs tâches de TAL (étiquetage, analyse syntaxique, catégorisation de documents, etc.), et présente l'avantage de diminuer considérablement le volume nécessaire de données annotées pour l'entraînement (Tang *et al.*, 2002). La problématique centrale des travaux dans ce domaine concerne le mode de sélection des items pertinents à soumettre à l'expert. Selon une étude récente dans la communauté du TAL (Tomanek et Olsson, 2009), ce type de méthode semble pourtant sous-utilisé, bien qu'il soit adapté à un grand nombre de situations.

Si l'apprentissage actif vise exclusivement la réduction des efforts d'annotation, il ne semble pas avoir été spécifiquement considéré comme un moyen d'adaptation à des données particulières, c'est-à-dire comme source d'informations modificatrices d'un modèle par défaut. Si la mise en place d'une telle procédure est bien entendu coûteuse (notamment en développement d'une interface dédiée pour l'interrogation de l'expert), elle repose sur le pari qu'elle peut intéresser les utilisateurs d'un analyseur qui cherchent à améliorer un traitement dans une situation particulière, et ne se contentent pas d'un système prêt à l'emploi. Ce point, mis en avant par Habert *et al.* (1997), qui précisent que toute annotation automatique doit généralement être accompagnée d'un travail manuel en amont ou en aval, est peut-être à relativiser, notamment au vu des volumes qu'ont à traiter les analyseurs dans les travaux actuels. Mais de tels efforts vont également dans le sens d'une meilleure observation des mécanismes de l'analyse syntaxique statistique, et pourront avoir des retombées dans l'amélioration de l'interaction avec la communauté du TAL plus orientée vers les questionnements linguistiques.

Un dernier point envisagé pour ces travaux de développement de l'analyseur est de considérer cette fois son utilisation, et d'adapter ainsi son fonctionnement aux exigences particulières de l'application visée. Puisque les analyseurs syntaxiques robustes font maintenant partie intégrante de la boîte à outils du TAL, leurs utilisations se diversifient grandement. Par contre, cela ne veut pas dire que les résultats attendus sont identiques dans ces différents contextes. Que ce soit en terme des types de relations exploitées, ou de la finesse des distinctions souhaitées, des variations sont identifiables. Par exemple, l'extraction de candidats-termes, la mesure de la complexité syntaxique ou l'analyse distributionnelle peuvent avoir des degrés d'exigence et de couverture variable. Il est donc nécessaire d'envisager l'impact des différents paramétrages et des injections de ressources au regard des besoins finaux.

Les travaux d'Assaf Urieli se situent donc au cœur de la problématique actuelle du TAL quantitatif. Ils posent clairement la nécessité de s'approprier une méthode quantitative par-

ticulière pour pouvoir envisager précisément son utilisation et permettre une intervention au coeur du programme pour pouvoir le positionner au sein des problématiques de la discipline. Cette maîtrise du système (ici la classification par entropie maximale) est également très importante pour envisager la réutilisation du classifieur dans d'autres contextes comme on l'a vu pour la classification des rapports d'incidents et comme on va le verra au chapitre suivant.

Les quelques travaux présentés ici soulèvent un ensemble de questions autour de l'utilisation des méthodes par apprentissage. Dans le prochain et dernier chapitre, je vais tenter de problématiser un ensemble de points et de dégager des interrogations et des pistes relatives à l'utilisation de ces méthodes face à des problématiques linguistiques.

Chapitre 8

Articulations de la linguistique et des méthodes d'apprentissage et de fouille de données

Après avoir montré au chapitre précédent les différentes utilisations envisageables des techniques par apprentissage, que ce soit pour explorer des données ou pour répondre à des besoins plus appliqués, je souhaite ici soulever un ensemble de points sur le rôle de ces techniques dans le TAL actuel.

Je vais commencer par un panorama plus épistémologique du développement et de l'évolution de ces techniques, en dégageant un ensemble de situations que j'estime problématiques. La question principale que j'aborde ici est celle de la complexité et de l'opacité de ces techniques, qui créent une démarcation au sein du TAL au sens large, et entre les acteurs qui les développent et les utilisent d'une part, et la famille des linguistes travaillant sur des données d'autre part.

J'aborderai ensuite les avantages que ces méthodes présentent pour un ensemble de questions liées aux données langagières, avant de présenter un exemple récent de leur utilisation pour une tâche spécifique (l'attribution d'auteur à un texte), sur lequel je m'appuierai pour proposer un ensemble de réflexions et de pistes permettant de mieux envisager la place de ces méthodes dans l'outillage du linguiste.

8.1 La révolution de l'apprentissage automatique en TAL

Je commencerai par la désormais classique constatation de l'évolution majeure qu'a connue le TAL durant ces quinze ou vingt dernières années, et qui a vu une véritable explosion des méthodes par apprentissage qui en constituent actuellement le modèle dominant, voire exclusif, et qui s'étend sur l'ensemble des sous-domaines de la discipline.

8.1.1 Étendue et origines du changement

La plupart des nombreux auteurs qui se sont penchés sur l'évolution de la discipline n'hésitent guère à parler de changement de paradigme, voire de révolution. L'argument principal consiste à remarquer que l'entrée en scène des méthodes par apprentissage a modifié les composants essentiels de la culture scientifique disciplinaire, que ce soit les approches concrètes

pour aborder un problème, le rapport aux données, la notion même de modèle, les liens avec les connaissances linguistiques, les pratiques publicatoires, et bien entendu la formation des étudiants et le bagage de connaissances attendues pour un jeune chercheur de la discipline.

On peut dater les premières secousses de la révolution au début des années 1990, en faisant confiance à Church (2011) qui en est un des principaux artisans. Ce qui ne semblait initialement qu'un retour bienvenu du TAL à des méthodes empiriques, préférant le recours à des données massives au développement de modèles ne pouvant généralement traiter que des exemples-jouets, a pris une ampleur que lui-même juge actuellement disproportionnée.

La preuve la plus tangible de cette (r)évolution concerne les publications en TAL, Church estime avec d'autres [1] que le taux d'articles utilisant des méthodes statistiques (au sens large, mais généralement il s'agit de méthodes par apprentissage) est passé de moins de 30% au début des années 1990 à plus de 90% à la fin des années 2000. Abney (2011) ajoute que le type même des publications a également totalement changé : si au début de cette période la publication prototypique consistait en la présentation d'une méthode de traitement exemplifiée sur deux ou trois cas, actuellement il s'agit exclusivement de la présentation d'une manipulation expérimentale sur une collection de test plus ou moins standardisée, cette manipulation étant rigoureusement évaluée mais bien souvent dépourvue d'exemples.

Cette évolution du TAL se place selon Church (2011) dans un contexte plus large, en la considérant comme une des conséquences de ce que l'on appelle parfois un *hiver* de l'Intelligence Artificielle. L'enthousiasme qu'a suscité cette discipline au long de son histoire a, à plusieurs reprises, entraîné d'énormes déceptions par des promesses non tenues. La fin des années 1980 correspond à une telle phase de désillusion, avec le déclin des systèmes experts dont on avait prétendu qu'ils allaient pouvoir automatiser une grande partie du raisonnement humain en situation complexe sur la base de règles logiques décrivant le fonctionnement d'un domaine d'application. Ces périodes hivernales se traduisent classiquement par l'abandon du financement de programmes scientifiques de grande ampleur, comme cela avait été le cas à la suite du rapport ALPAC qui avait signé un coup d'arrêt aux travaux sur la traduction automatique en 1966.

De fait, c'est sans doute également par le biais de la traduction automatique que le changement de paradigme est, entre autres, arrivé au centre du TAL (après avoir été pleinement développé dans un domaine connexe, la reconnaissance de la parole). La traduction statistique, basée sur des corpus alignés, s'est rapidement imposée par son efficacité et son faible coût de développement, lorsqu'on la compare aux modèles à base de règles. Les principaux autres champs parmi les premiers concernés furent les techniques d'étiquetage de texte, et parmi elles l'analyse syntaxique automatique.

La figure 8.1 de la page 200 présente une frise chronologique de ces évolutions en positionnant quelques événements majeurs dans la discipline. J'y ai représenté d'un côté les inventions des principales techniques génériques d'apprentissage, les créations des outils les plus couramment utilisés pour ces tâches et leurs utilisations notables en TAL. Y sont également placés sur l'axe chronologique les grands événements indiquant l'impact sur la communauté des méthodes par apprentissage : quelques conférences dédiées, les débuts des grandes campagnes d'évaluation, et l'ouvrage de référence de Manning et Schütze (1999). La courbe superposée reprend l'estimation évoquée plus haut de la proportion des publications de l'ACL qui font usage de méthodes quantitatives (l'origine est sur l'axe temporel et le trait horizontal

1. Notamment l'étude du corpus des publications de l'*Association for Computational Linguistics* réalisée par Hall *et al.* (2008).

supérieur correspond à un niveau de 100%). Les notions et les techniques particulières sont évoquées au cours de ce chapitre ou du précédent. On peut ainsi y voir la progression de ces méthodes, et comment elles ont profondément et rapidement influencé la discipline dans ses pratiques concrètes et institutionnelles.

Il faut noter enfin que ce schéma partiel ne traduit pas la grande multiplicité des applications des méthodes par apprentissage, qui sont devenues l'approche majoritaire, voire exclusive, des différents travaux de recherche et d'application en TAL.

8.1.2 Redéfinition des rapports entre les disciplines

Cette évolution rapide et radicale a eu plusieurs impacts sur la façon d'aborder les différents pans du TAL, et a contribué à creuser plusieurs fossés entre eux. Deux dimensions principales sont à noter : la première concerne la distinction entre les volets théoriques et appliqués du TAL, et la seconde entre la linguistique et l'informatique.

8.1.2.1 Science ou ingénierie ?

La première distinction pose toujours des problèmes de terminologie en français : si l'anglais a su se doter de deux termes bien distincts, *computational linguistics* et *natural language processing*, je n'ai toujours pas l'impression que des distinctions équivalentes soient véhiculées par une opposition entre *TAL* et *linguistique informatique*, voir à ce sujet (Cori et Léon, 2002). Selon Kay (2011), la distinction tend à faire pencher le *natural language processing* vers l'ingénierie, et c'est sur lui que repose l'essentiel de la transformation de la discipline. En effet, la révolution de l'apprentissage automatique a surtout apporté des solutions pratiques, efficaces et peu coûteuses à des problèmes concrets du traitement du langage. Si, comme lui, on considère que la *computational linguistics* est une véritable branche de la linguistique, qui se distingue des autres par ses moyens d'investigation empruntés aux techniques et aux modèles informatiques, la montée en puissance des méthodes par apprentissage a effectivement creusé l'écart avec le volet précédent.

La première question soulevée concerne en effet le gain en termes de connaissances sur le fonctionnement de la langue qu'apportent des approches souvent très opaques dans leur fonctionnement, et dont les mécanismes intrinsèques sont rarement accessibles, même si leur efficacité est aisément démontrée. L'autre distinction concerne la place accordée aux théories et connaissances linguistiques dans les systèmes : si c'est une simple application qui est visée, il est malheureusement acquis comme le rappelle entre autres Johnson (2011) que leur rôle est minime et décroissant. Johnson pense par contre que les interactions sont nettement plus prometteuses lorsque c'est la connaissance des mécanismes fondamentaux du langage (notamment son acquisition) qui est visée.

C'est bien souvent au moment de l'évaluation d'une méthode que la différence des objectifs est la plus visible, comme le note François Yvon :

> *Sur la question de l'évaluation toujours, on déplorera néanmoins que l'évaluation quantitative des modules de traitements, réalisée le plus souvent de manière indépendante des autres modules, soit devenu le « mètre étalon » unique des apprentis linguistiques. Les autres évaluations possibles des travaux liant apprentissage et traitement automatique des langues : leur adéquation avec une théorie linguistique ou avec des observations comportementales, leur apprenabilité, leur plausibilité cognitive, leur stabilité (qu'on pense à des tâches non-supervisées), l'interprétabi-*

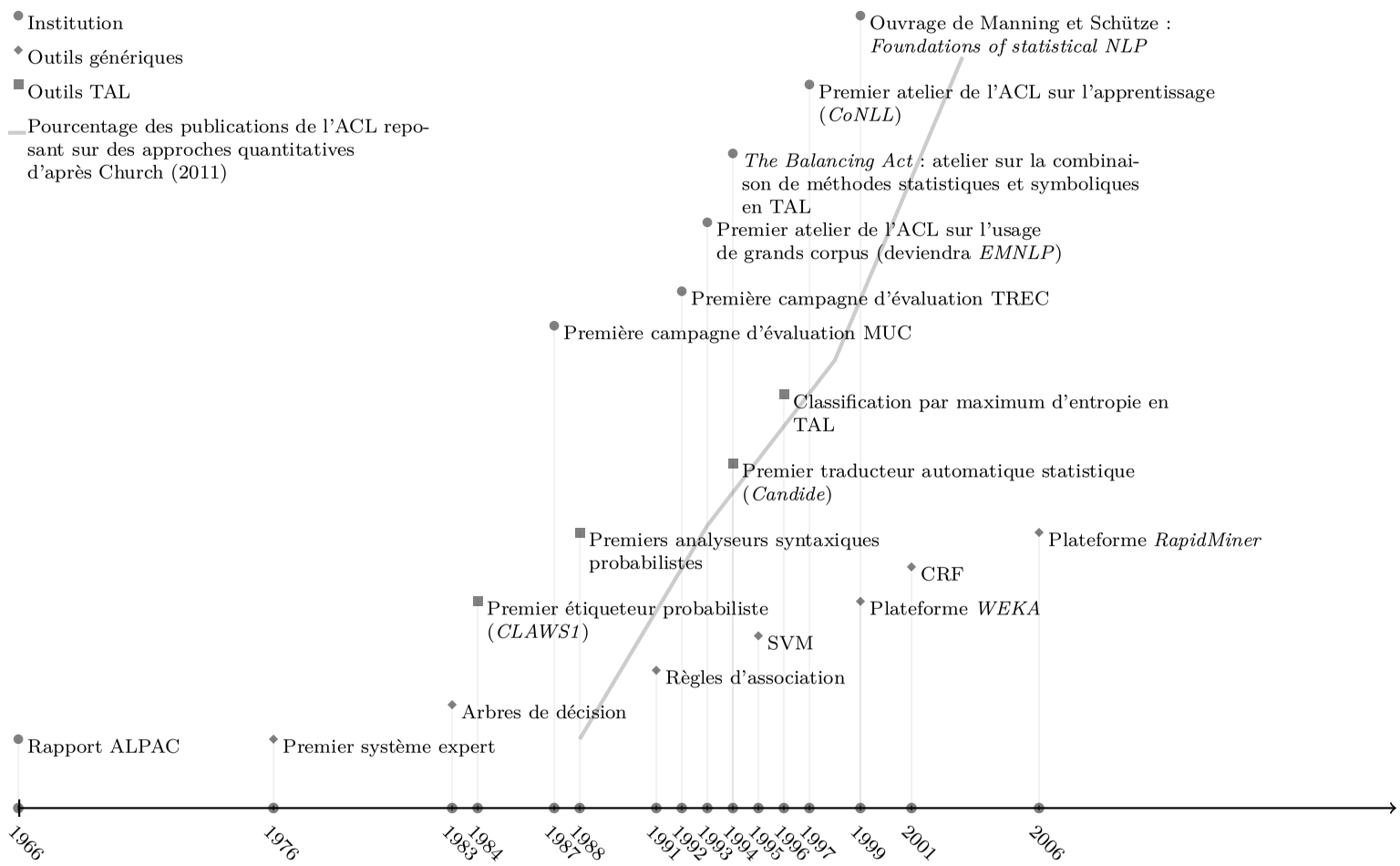

FIGURE 8.1 – Frise chronologique des méthodes par apprentissage en TAL

> *lité des modèles induits ou de leurs paramètres, sont ainsi progressivement passés au second plan des préoccupations.*
> (Yvon, 2006)

Mais la distinction entre les deux aspects de la discipline ne sont pas toujours aussi simples ni reconnus comme tels. Par exemple, Abney (2011) voit les approches par apprentissage comme un versant *expérimental* de l'investigation du fonctionnement de la langue. Les travaux actuels respectent en effet les procédures habituelles de ce type de travaux scientifiques, en précisant un protocole d'évaluation clair, une comparaison systématique des résultats avec des approches concurrentes sur des étalons préétablis, et l'utilisation de mesures statistiques pour évaluer le gain obtenu. Toutefois, si une expérimentation a pour but de tester la validité d'un modèle, celui-ci est dans les faits rarement issu de questionnements linguistiques ; les objectifs d'une comparaison expérimentale sont généralement applicatifs et motivés par la proposition d'une nouvelle méthode de manipulation des données.

Abney et Kay se distinguent donc par leur découpage de la discipline autour de ces questions de méthodes, et de ce fait autour des interactions (possibles ou réelles) avec la linguistique. Tous deux sont toutefois d'accord sur le fait que le développement de ces techniques a profondément modifié les rapports avec les linguistes sur des questions pourtant communes.

8.1.2.2 Place de la linguistique

On présente souvent l'âge d'or de la collaboration entre linguistes et informaticiens du TAL comme étant celui où les premiers proposaient des connaissances formalisées aux seconds qui les intégraient dans des systèmes automatisés. Idéalement, la connaissance linguistique ainsi injectée permettait d'accroître les performances des applications, mais cette interaction fournissait également aux linguistes une confrontation à large échelle de leurs propositions théoriques, et permettait un progrès dans les modèles. Cette image d'Epinal a sans doute été la plus justifiée dans le domaine de l'analyse syntaxique sur la base des différents formalismes issus de la grammaire générative. En tout cas, le linguiste pouvait à juste titre se considérer comme étant positionné au centre du système, même si les tâches étaient clairement (et sans doute à tort) séparées. Mais les techniques par apprentissage ont largement changé la donne, et attribué aux linguistes un rôle secondaire, quand elles ne l'ont pas tout simplement retiré du processus.

On se souvient longtemps après de la fameuse phrase de Jelinek [2] : « *Whenever I fire a linguist our system performance improves.* ». Appliquée au domaine de la reconnaissance de la parole, elle traduisait de façon caricaturale le rôle croissant des méthodes statistiques à la fin des années 1980. Si ce genre d'ostracisme est une vision faussée et extrême, il n'en est pas moins vrai que le rôle principal attribué au linguiste dans les méthodes par apprentissage se limite à l'annotation des données, que ce soit à des fins d'entraînement ou d'évaluation (Fabre, 2010).

Le constat est d'autant plus évident quand on regarde le contenu des cursus de TAL (aux États-Unis et au Royaume-Uni) et la place décroissante qu'y prennent les enseignements de linguistique générale. Autant Kay (2011) que Church (2011) le regrettent, et applaudissent les quelques cas où le retour semble s'opérer, pour l'essentiel en syntaxe. En France, les formations proposées par les départements d'informatique font évidemment une place centrale

2. Dans son allocution *Applying Information Theoretic Methods : Evaluation of Grammar Quality Workshop on Evaluation of NLP Systems* en 1988, commentée avec humour et remords dans (Jelinek, 2005).

aux méthodes par apprentissage en TAL (les deux domaines étant très souvent associés), et je ne crois pas que les enseignements de linguistiques soient en meilleure posture, mais l'ont-ils jamais été ? Pour les formations de TAL en sciences du langage, il me semble à l'inverse que les méthodes quantitatives prennent une place timidement croissante dans les cursus de linguistiques, suivant une évolution qui a déjà eu lieu dans les autres sciences humaines et sociales.

Plusieurs explications sont proposées pour cette séparation : Abney (2011) explique le rejet des méthodes probabilistes par les linguistes à cause de leur trop grande similarité avec le béhaviorisme, ou encore par le fait que les méthodes empiriques réapparaissaient du côté de la performance dans le sempiternel débat chomskyen. Il faut sans doute y ajouter que la maîtrise des modèles mathématiques complexes mis en œuvre dans les méthodes par apprentissage constitue en soi un obstacle à l'appropriation de ces méthodes, par opposition aux principes plus généraux comme celui de l'unification (même si la complexité des mécanismes de certains formalismes génératifs est à mon avis largement aussi imposante).

Plusieurs issues sont évoquées pour sortir de cette situation insatisfaisante de clivage entre deux communautés qui œuvrent en partie dans la même direction. La plus directe est celle qui met le linguiste dans la boucle du processus d'apprentissage, en l'impliquant par exemple dans des tâches d'apprentissage supervisé, en sollicitant son expertise face à des cas-limites, ou pour proposer des biais dans les pondérations calculées (Steedman, 2011). C'est ce que nous envisageons de faire autour de l'analyse syntaxique, comme indiqué en 7.3.4 (page 193).

Hajičová (2011) insiste quant à elle sur le rôle du linguiste dans la définition des traits descriptifs et de leur exploitation, c'est-à-dire dans l'injection de connaissances linguistiques là où bien souvent des informations de surface sont très majoritairement utilisées (voir section 8.4.3 page 217). Toutefois, les exemples qu'elle met en avant sont peu convaincants en ce qui concerne le recours à l'expertise linguistique dans l'utilisation des traits elle-même, par exemple en demandant au linguiste la taille ou la position relative du contexte pertinent pour une opération de désambiguïsation.

Kay (2011) et Hajičová (2011) rappellent que les connaissances linguistiques pointues sont les seules façons de faire progresser les systèmes au-delà de la zone des phénomènes suffisamment répétés, *i.e.* les fameux 80 premiers %, alors que la loi de Zipf nous a depuis longtemps appris qu'un grand nombre de phénomènes très rares constituent la partie manquante de la couverture de tout système de TAL. Puisque l'approche probabiliste par des corpus massifs tombe la tête la première dans ce piège, seule la connaissance théorique de ces phénomènes sporadiques permettra leur traitement efficace. On peut rapprocher cette question des gains minimes (et pas toujours mesurables) qu'apportent des connaissances linguistiques aux performances d'un système par apprentissage, qui ne justifient pas toujours le coût de leur développement. J'aborde cette question sur un exemple en 8.3.

Comme d'autres linguistes confrontés directement à cette rupture, Fabre (2010) et Habert (2004) plaident à juste titre pour un rapprochement plus actif de la part des linguiste eux-mêmes, en incitant à une appropriation d'une partie des techniques initialement incriminées, et notamment à travers la formation des étudiants en sciences du langage.

Dans tous les cas Church (2011) prédit un retour cyclique du balancier vers des travaux de TAL plus théoriques et moins empiriques lors de la prochaine décennie, avec dans ce cadre une remontée en puissance de la linguistique sur la scène du TAL. Il reste à voir de quel côté viendra ce retour, et si la linguistique descriptive saura profiter efficacement de cette opportunité.

8.1.3 Évolution du rapport aux données

En attendant ce retour à un plus juste partage des connaissances et une pratique commune, force est de constater que le fossé actuel est également présent sur le plan du rapport aux données.

Levin (2011) évoque un curieux paradoxe, en rappelant que la révolution s'est justement imposée sur la base d'une revendication d'un retour massif aux données par opposition aux modèles abstraits. Les acteurs de l'apprentissage automatique en TAL sont en effet les plus grands consommateurs de données (*There's no data like more data*), mais semblent également être ceux qui évitent le plus la confrontation directe avec celles-ci :

> *Proponents of statistical approaches love large data sets, but most seem to be afraid to touch the data with their bare hands, prefering instead to handle it with models and automatic scoring metrics. [...] The people who say they love data the most seem to be the most afraid of looking at it. [...]*
> *Perhaps the issue is strong faith in statistical methods, which is sometimes justified. However, it is more likely that researchers see languages as black boxes because they lack meta-linguistic knowledge about how languages are structured.*
> (Levin, 2011)

Il est en effet étonnant de voir le peu de regards posés sur les données utilisées dans les travaux utilisant massivement l'apprentissage automatique. On peut remarquer notamment l'absence généralisée d'exemples dans nombre de publications, et le seul recours à des étalons prédéfinis et des scores globaux pour mesurer l'efficacité d'un traitement. Sur une note plus anecdotique, lors de la campagne PAN 2011 (voir section 8.3, page 210), une récente compétition sur l'attribution d'auteur, nous étions (avec mes co-auteurs) les seuls à montrer dans nos transparents des extraits des textes ciblés, et à avoir observé sur quelques cas le résultat de nos traitements, là où tous les autres participants se focalisaient exclusivement sur des scores comparés de rappel et de précision.

Il semblerait également que l'observation à l'œil nu des données langagières manipulées soit absente des processus de mise au point des méthodes elles-mêmes. Cela semble être le cas dans d'autres disciplines comme la recherche d'information et ses dérivés. Le point commun est bien entendu la quantité impressionnante de données manipulées qui décourage ce type d'observation, et la présence centrale des bancs de test qui jouent le rôle des instruments de navigation et détournent le regard du pilote du paysage. Comme on peut le voir dans la frise chronologique de la figure 8.1, les campagnes d'évaluation ont naturellement accompagné la montée en puissance des méthodes par apprentissage. Il semblerait donc étonnant que ce soit, comme le propose Levin, la simple méconnaissance linguistique qui soit à l'origine de ce comportement.

De plus, la situation est d'autant plus grave à mes yeux que l'observation des mécanismes plus précis appliqués aux données est elle aussi absente, rendue pratiquement impossible par les modèles purement probabilistes et numériques qui n'offrent qu'une interprétabilité très limitée (voire inexistante).

8.1.4 Évolution des descripteurs pour les applications

Comme je l'ai décrit plus en détails en section 7.1 (page 175), toute approche basée sur l'apprentissage automatique procède par le calcul de traits descripteurs associés à l'unité linguistique traitée (mot, segment ou texte). C'est sur la base de ces descripteurs que sont

apprises les régularités sur lesquelles repose exclusivement le processus. Dès lors, le choix de ces descripteurs, et de leur mode de calcul constitue une étape vitale pour laquelle il est possible d'envisager une très grande variété de caractéristiques, parmi lesquelles on peut trouver des propriétés linguistiques complexes.

Toutefois, il semble qu'une place croissante et tendant à devenir exclusive soit accordée dans les travaux actuels de TAL à des descripteurs de surface, au détriment des informations plus riches que permettent d'obtenir les différentes techniques d'analyse et les ressources langagières génériques. L'exemple central de ce phénomène est la prédominance des n-grammes (de caractères ou de mots) dans les tâches de classification de textes, en lieu et place par exemple des informations syntaxiques ou des structures plus complexes.

Par exemple, Gamon (2004) aborde une tâche désormais classique, la détection des sentiments et des opinions dans les avis de clients sur un produit en utilisant comme descripteurs les habituels n-grammes de mots. Toutefois, il étudie également le rôle de traits linguistiques plus sophistiqués (séquences d'étiquettes morpho-syntaxiques, temps verbaux, taille des constituants syntaxiques, structures des syntagmes, etc.) pour en évaluer la valeur ajoutée. On voit dans sa remarque à la suite de ces calculs qu'il s'agissait bien d'une tentative non-standard dans le domaine :

> *A result that came as a surprise to us is the fact that the presence of very absract linguistic analysis features based on constituent structure and semantic dependency graphs improves the performance of the classifiers. [...] While the improvement in practice may be too small to warrant the overhead of linguistic analysis, it is very interesting from a linguistic point of view that even in a domain as noisy as this one, there seem to be robust stylistic and linguistic correlates with sentiment.*
> (Gamon, 2004)

Si l'on peut se réjouir d'une telle conclusion, on prend bien là la mesure de la distance entre les approches de TAL traditionnelles et ce genre d'applications (qui forment le gros des travaux actuels). On verra plus loin que c'est le cas dans de nombreuses autres tâches, pour lesquelles les traits linguistiques complexes ont depuis longtemps été abandonnés, comme le note également Witten pour la fouille de textes :

> *But, in fact, most text mining efforts consciously shun the deeper, cognitive, aspects of classic natural language processing in favor of shallower techniques more akin to those used in practical information retrieval.*
> (Witten, 2005)

Les raisons de cet abandon sont évidentes : la première est le faible coût de calcul (une simple segmentation en mots et une lemmatisation tout au plus, voire aucun traitement dans le cas des n-grammes de caractères), à laquelle vient s'ajouter l'indépendance vis-à-vis de la langue et des données traitées qui permet une réutilisation économique des outils.

Comme indiqué par Witten, cet état de fait est similaire à celui de la recherche d'information, champ applicatif pour lequel la recherche d'une valeur ajoutée des techniques linguistiques est depuis longtemps abandonnée, comme l'indiquait déjà Sparck-Jones (1999).

> *Thus claims that linguistic analysis is needed for other indexing purposes than the provision of structured index descriptions on which I have concentrated have not yet been substantiated. [...] It is not clear, either, that NLP is required for some tasks that are closely related to ordinary retrieval.*
> (Sparck-Jones, 1999)

Le débat autour de ces différences d'approche du matériau langagier n'est pas un rejet de la pertinence des phénomènes linguistiques envisagés par les approches plus linguistiques. L'argument scientifique principal en faveur des descripteurs de surface est qu'ils permettent de capter des phénomènes complexes en compensant leur pauvreté par le volume, et que les régularités qui émergent d'une approche quantitative rejoignent une grande partie des connaissances linguistiques injectées par les analyseurs ou les ressources génériques. C'est ce type de débat qui est à l'œuvre autour des analyseurs syntaxiques statistiques par opposition aux modèles génératifs, par exemple autour des relations syntaxiques à longue distance et l'impossibilité démontrée par Chomsky de leur traitement par des méthodes à états finis (Church, 2011).

Toutefois, comme on le verra plus loin, plusieurs expériences montrent au contraire que les descripteurs linguistiques peuvent tout à fait être utilisés avec profit pour des tâches qui les ont ignorés ou rejetés. Dans certaines situations, ils sont même plus efficaces que les descripteurs de surface, puisqu'ils injectent dans une application des informations déjà disponibles. Ils peuvent donc se montrer avantageux dans les situations (et elles existent encore !) où la masse disponible n'est pas suffisante, et pour dépasser la fameuse limite que la distribution zipfienne impose, en traitant ces phénomènes trop rares pour être suffisamment bien repérés.

8.1.5 Exemple de l'évolution des approches : l'analyse des références bibliographiques

Je terminerai ce tour d'horizon général par un exemple plus local, qui traduit bien les évolutions de la discipline dans la façon d'aborder les tâches mineures souvent rencontrées pour pouvoir exploiter des données. L'exemple concerne ici celui de l'analyse des références bibliographiques qui fut une des étapes initiales de l'analyse linguistique des contextes des appels de citation dans le cadre du projet Rhecitas (voir section 5.2.3, page 126). Afin de pouvoir identifier dans le corps du texte ces appels de citation, il était en effet nécessaire d'extraire de la bibliographie (généralement aisément repérable à la fin d'un article scientifique) les éléments de chaque item, notamment les noms et prénoms des auteurs, ainsi que l'année de publication (ce sont généralement ces éléments qui constituent les appels de citation). Avec nos partenaires de l'INIST (Claire François et Dominique Besagni), nous avons appliqué la méthode qui nous paraissait la plus naturelle, à savoir le développement de patrons génériques à base d'expressions régulières, ainsi que différentes heuristiques traduisant notre connaissance des normes en vigueur (et de leurs variations) pour l'écriture de références bibliographiques. C'est d'ailleurs ce genre de pratiques qui avaient été développées par le passé, notamment pour des outils comme Paracite [3].

Quelques années plus tard, à l'occasion du projet CAAS, je découvrais que Patrice Bellot et ses collaborateurs travaillaient au développement d'un outil similaire dans le cadre du projet BILBO (Kim *et al.*, 2011). Cette fois par contre, c'est une méthode par apprentissage qu'ils ont utilisée : en lieu et place du développement de grammaires locales, ils ont annoté manuellement un échantillon de références (en délimitant précisément chaque champ : nom, prénom, année, titre, etc.) et appliqué un modèle probabiliste, en l'occurrence les CRF (*conditional random fields*). Cette méthode d'apprentissage relativement récente est une extension des chaînes de Markov qui attribue de façon probabiliste une catégorie à chaque élément d'une séquence en fonction de ses propres caractéristiques et de la nature des éléments qui l'entourent. Ce

3. `http://paracite.eprints.org`

genre de méthode est désormais couramment utilisée et vue comme très prometteuse pour plusieurs tâches similaires en TAL, comme l'étiquetage morphosyntaxique. Les descripteurs utilisés ici pour décider du statut de chaque sous-chaîne d'une référence bibliographique sont des éléments typographiques (présence de capitales, de chiffres, de parenthèses, etc.).

On voit bien le changement culturel à l'œuvre ici, puisqu'il touche désormais un ensemble très important de tâches. Au lieu de définir et projeter des mécanismes qui expriment directement les indices et les décisions à prendre en fonction de ceux-ci (par exemple, le repérage des virgules pour décider de la position relative du nom et du prénom de l'auteur), ce sont maintenant les régularités observées sur des données annotées qui permettent la prise de décision par le programme. Là où la collaboration avec des spécialistes des normes de notation bibliographique se traduisait par leur implication dans le processus de définition des règles, elle est maintenant décorrélée du produit final et se limite à l'annotation du corpus d'entraînement (on peut même se demander s'il s'agit encore d'une véritable collaboration).

On pourrait être tenté de faire un parallèle entre cette évolution et celle qui a traversé l'IA des années 1980 et fait émerger l'ingénierie des connaissances. Les premiers systèmes experts, programmes conçus pour résoudre un problème et prendre une décision face à une situation donnée, étaient alimentés par des règles logiques proposées directement par un expert du domaine. Cette phase d'acquisition des connaissances était initialement longue et fastidieuse, peu naturelle pour l'expert en question qui devait expliciter les mécanismes de son raisonnement là où son habitude était de pratiquer sur des cas concrets. Un ensemble de techniques a donc été développé pour faciliter cette opération, notamment en présentant à l'expert des situations typiques à résoudre, et en lui faisant éventuellement commenter sa décision [4].

Toutefois, il ne m'apparaît pas évident que la tâche d'analyse d'une référence bibliographique soit d'une complexité similaire à celles qui ont fait les beaux jours des systèmes experts (diagnostic médical, paramétrage d'une chaîne de production industrielle, etc.), ni que les descripteurs à intégrer soient d'une quantité difficilement contrôlable ou difficiles à identifier.

De fait Kim *et al.* (2011) envisagent d'ajouter des règles spécifiques en post-traitement pour compléter et corriger leur analyseur, ce qui correspond également à une tendance courante dans les méthodes par apprentissage, et constitue une des façons de concilier les différentes philosophies.

8.1.6 Vers un équilibrage des méthodes et des cultures

Malgré ce constat de clivage profond, de nombreux signes indiquent qu'une évolution est en cours, et vise à rapprocher les différentes tendances en TAL, en intégrant de façon plus évidente les connaissances linguistiques aux méthodes par apprentissage.

La première façon de l'envisager est de développer des méthodes *hybrides*, comme indiqué dans l'exemple précédent. Un tel mélange de deux façons d'aborder une problématique de TAL (généralement appliquée) consiste à faire collaborer une méthode par apprentissage avec un traitement classique (à base de règles pour simplifier), que ce soit en entrant des informations riches dans le système statistique, en appliquant des traitements successifs, ou en les faisant travailler séparément avec une union de leurs résultats (par un système de vote par exemple). Cette notion d'hybridation était déjà présente il y a plusieurs années, traduite notamment par un atelier spécifique sur la question en 1994 (Klavans et Resnik, 1996), et continue à susciter

4. Par la suite, de telles connaissances ont été essentiellement recherchées dans les textes spécialisés et les écrits des mêmes experts. Ces derniers ne sont alors impliqués qu'à la fin du processus pour valider les résultats.

de l'intérêt [5].

Cette voix médiane entre les deux approches est d'ailleurs vue avec enthousiasme par les acteurs de l'apprentissage automatique, comme l'indique par exemple Isabelle Tellier dans la préface du numéro spécial qu'a consacré la revue TAL à la question de l'apprentissage automatique :

> *Le TAL, qui n'a jamais non plus renoncé à l'idéal d'universalité des sciences du langage, a tout à gagner à cette nouvelle maturité. Mieux, l'apprentissage automatique pourrait lui permettre en quelque sorte de se réconcilier avec lui-même : la rupture historique, en son sein, entre « grammaires formelles » et théorie de l'information, qui remonte aux controverses entre Chomsky et Harris, a de moins en moins lieu d'être, quand on regarde de près les travaux actuels qui combinent les deux approches. Il n'y a plus vraiment de contradictions à construire manuellement des ressources ou des modèles formels et à les exploiter dans un programme d'apprentissage automatique à partir de données. Les acquis de l'apprentissage automatique doivent désormais faire partie du bagage de base de tout bon praticien du TAL.*
>
> (Tellier, 2009)

Il reste cependant à mon avis à franchir le pas de l'autre côté de la frontière, et voir quel bénéfice peuvent apporter ces méthodes et cette façon de travailler aux partisans des connaissances linguistiques plus explicites. Pour l'instant, cet aspect ne semble concerner que les avantages que présentent les méthodes quantitatives « classiques », comme le résume Martin Kay :

> *Fortunately, this chapter in the history of our field will be short lived. Already, the realization is growing that languages have morphologies, that adjacency in the string is the wrong domain of locality for many linguistic purposes, that sentences have recursive structures, that there is a difference between what you say and the language you say it in. For their part, linguists are coming to appreciate that, while statistics explain nothing by themselves, they can cast light on what needs to be explained and, perhaps, where to start looking for the explanation.*
>
> (Kay, 2011)

J'estime que les derniers points évoqués par Kay ci-dessus sont important à développer de l'autre côté (donc sur la rive linguistique), et c'est précisément ce que j'ai commencé à faire en appliquant des techniques quantitatives à des données langagières complexes.

Je vais maintenant tâcher de préciser le contexte local, en précisant les motivations qui nous ont amené progressivement à utiliser les méthodes par apprentissage.

8.2 Motivations pour l'utilisation des méthodes par apprentissage

Avant tout, il est important d'affronter de face le fait que l'évolution globale ne peut être ignorée, indépendamment des choix et des positions scientifiques personnelles ou à l'échelle

5. Un atelier sur cette question est proposé à la conférence EACL en 2012.

d'une équipe. La montée en puissance des approches par apprentissage peut être vue comme une mode par certains, mais elles se sont imposées avec suffisamment de force et suffisamment entraîné la conviction pour qu'elles soient perçues comme un passage obligé de tout travail en TAL.

Il était donc normal, et cohérent avec le rôle que je me suis donné dans mon environnement professionnel immédiat, d'accompagner et de participer aussi pleinement que possible à ce changement culturel. Je peux donc dégager une série de motivations qui m'ont poussé à travailler spécifiquement avec ces méthodes.

8.2.1 Facilité et rapidité pour le développement des applications

Si les méthodes par apprentissage ont connu un tel succès sur le plan des applications, c'est essentiellement par la rapidité et la facilité (relative) de leur utilisation face à un besoin concret. L'exemple sans doute le plus central est celui des analyseurs robustes : depuis longtemps nous sommes habitués à utiliser des systèmes conçus par apprentissage pour l'étiquetage morphosyntaxique (comme le célèbre mais vieillissant TreeTagger, une des rares ressources disponibles librement pour le français, fonctionnant comme son nom l'indique à l'aide d'arbres de décision).

Du côté de l'analyse syntaxique, la situation n'était pas si monopolistique, et jusqu'à il y a peu la plupart des analyseurs du français se basaient sur des grammaires ou des règles explicites. C'est bien entendu le cas de l'analyseur Syntex, qui fut développé à l'ERSS et largement utilisé par mes collègues et moi-même dans un grand nombre de situations diverses (voir section 3.3.1, page 65).

Malheureusement, le départ de Didier Bourigault, et le choix stratégique de dépôt conjoint d'un brevet avec la société Synomia a entraîné l'impossibilité d'utiliser cet outil pour nos besoins et ceux de nos autres partenaires. La question s'est donc posée de trouver un outil de remplacement. Nos exigences avaient malheureusement pris de l'ampleur avec Syntex : en plus d'un analyseur robuste et très efficace, nous avions en plus pris l'habitude de le configurer (presque) à notre guise, par exemple en pré-traitant certains corpus, et bien entendu en interagissant directement et quotidiennement avec son concepteur. Autrement dit, nos besoins ne pouvaient être tous satisfaits par l'utilisation d'un outil prêt-à-l'emploi, non paramétrable et opaque dans son fonctionnement.

Le développement de Syntex était le fruit d'un travail de plusieurs années, avec la mise au point minutieuse de règles spécifiques pour chaque type de relation syntaxique et de leurs différentes articulations (Bourigault, 2007). Nous ne disposions ni de l'énergie ni du temps nécessaire pour nous relancer dans un développement de ce type.

Entre temps, le monde, comme on l'a vu, avait largement opté pour le développement d'analyseurs statistiques, construit par apprentissage sur des corpus annotés comme le French Treebank (Abeillé *et al.*, 2003).

La décision fut prise par l'arrivée dans notre équipe de Assaf Urieli, qui souhaitait effectuer, pour son doctorat sous mon encadrement, le développement et l'approfondissement d'un tel outil (voir 7.3.4, page 193). Si plusieurs outils existant étaient déjà disponibles ou en cours d'adaptation, notamment grâce à l'équipe Alpage (Crabbé et Candito, 2008), nous souhaitions en effet « avoir la main » sur l'outil complet, afin notamment d'envisager son adaptation à différents contextes et utilisations (intégration de ressources langagières, ciblage de certaines relations pour des besoins spécifiques, etc.). Le choix d'Assaf Urieli s'est porté sur les classifieurs par entropie maximale, basé sur des indices de surface pour calculer les

relations de dépendances syntaxiques, et l'outil Talismane devrait voir prochainement le jour. Le temps de développement, même s'il n'est bien entendu pas négligeable, n'est toutefois en aucun cas comparable avec celui de Syntex. Quant à la qualité des résultats, les différentes évaluations ciblées que nous en ferons nous donneront à terme une estimation de la perte entraînée par ce changement de méthode.

Un dernier aspect, peut-être plus anecdotique, concerne le rôle que jouent les campagnes d'évaluation organisées au sein du TAL pour un nombre croissant de tâches. Comme on le verra dans l'exemple de l'attribution d'auteur (voir 8.3), la temporalité de ces campagnes, et notamment le faible temps disponibles entre la mise à disposition des exemples (en fait, des données d'entraînement) et la soumission des résultats ne laisse de fait pas la place pour des développements relevant d'une autre façon de faire.

8.2.2 Enthousiasme des étudiants

La prédominance des méthodes par apprentissage est très clairement perçue comme telle par nos étudiants, même si nous ne mettons pas particulièrement l'accent sur celles-ci dans notre formation en TAL à Toulouse. Que ce soit par une forme compréhensible de fascination par rapport à ces méthodes (qui semblent de loin, comme par magie, résoudre seules les problèmes et faire émerger des réponses à des questions complexes), ou par simple volonté de ne pas rester en marge, ce sont souvent nos étudiants qui ont décidé par eux-mêmes de se lancer dans l'aventure de l'apprentissage en TAL. En plus d'Assaf Urieli mentionné ci-dessus, ce choix a également été fait par Marion Laignelet, qui a collaboré avec l'équipe du GREYC à Caen autour de l'exploitation d'un grand nombre de marqueurs linguistiques pour identifier les segments de textes obsolescents (voir section 7.3.1, page 187).

Par la suite, et notamment pour les travaux des étudiants arrivés après, le pas était déjà franchi, et de plus en plus la décision d'employer ce genre de méthode est désormais à la fois naturelle et collective (en tout cas jusqu'ici, pour les doctorats de Nikola Tulechki et Simon Leva).

8.2.3 Simplicité pour l'analyse des données

J'ai présenté à travers les précédents chapitres les besoins que les différents travaux toulousains en terme d'annotation de corpus soulevaient quand il s'agissait d'en analyser les résultats. Si les méthodes statistiques classiques présentées au chapitre 6 sont le choix classique pour ce genre de travail, elles peuvent également être en partie remplacées par des méthodes de fouille de données. On a vu au chapitre précédent (section 7.2, page 182) comment des résultats similaires peuvent être obtenus, généralement plus rapidement et plus facilement avec des méthodes de ce type. Le parallélisme entre les deux types de techniques est évident, même si les méthodes par apprentissage et par fouille n'offrent pas les mêmes garanties sur les résultats ni le même contrôle sur les détails des méthodes (j'y reviens plus en détails en section 8.4.1, page 215).

8.2.4 Un dispositif expérimental pour de nouvelles méthodes et ressources

La dernière motivation est à mon avis la meilleure, car elle tend à chercher des réponses aux différentes questions soulevées dans les débats que j'ai tenté de résumer ici. Elle concerne

la vision (celle d'Abney) des méthodes par apprentissage comme un dispositif expérimental, capable au final (si l'on insiste suffisamment) de nous apprendre quelque chose, sinon sur le langage, du moins sur certains phénomènes ou certaines façons de les aborder.

Les explorations du langage par des techniques de TAL sur corpus tendent en effet à produire un ensemble de données secondaires, qui peuvent constituer autant de ressources exploitables (en tant que descripteurs) par un système par apprentissage. C'est par exemple le cas des classes lexicales induites par l'analyse distributionnelle, ou encore des différents marqueurs (syntaxiques ou discursifs) étudiés et modélisés. Ces données parfois éparses et grossières, souvent difficiles à cerner, peuvent tout à fait être exploitées (pratiquement telles quelles), et si possible étudiées dans leur participation à une tâche appliquée comme la classification de textes. Il est ainsi possible à la fois de tester leur pertinence pour la tâche, mais aussi leur comportement en tant que marqueurs descriptifs, pour peu que le système d'apprentissage utilisé se prête à une lecture plus ou moins directe du modèle induit.

On retrouve dans ce type d'approche l'héritage de Biber (1988), en ce sens que les systèmes d'apprentissage (pour la plupart) permettent de positionner en entrée un grand nombre de descripteurs disparates, et qu'un des résultats secondaires de la phase d'entraînement est justement une évaluation de leur pertinence individuelle et collective. Comme on le verra, nous avons ainsi participé à une tâche de classification automatique de documents en injectant une grande variété de traits linguistiques, parfois très complexes, en déléguant au système par apprentissage le soin de démêler les écheveaux (voir section suivante).

Cette utilisation opportuniste des méthodes par apprentissage s'accompagne d'une investigation plus précise du rôle exact des techniques linguistiques « riches » dans ce genre d'application, s'opposant par là à l'appauvrissement croissant des descripteurs. Peut-être s'agit-il d'un combat perdu d'avance, mais je reste persuadé avec d'autres que les connaissances de la linguistique n'ont pas dit leur dernier mot face aux trigrammes de caractères. Ce débat est notamment toujours actuel dans le domaine de la recherche d'information, malgré les conclusions négatives de Sparck-Jones, voir par exemple Moreau *et al.* (2007).

Dans tous les cas, ce genre de rapprochement culturel est bénéfique : bien au-delà de la recherche d'une meilleure performance pour une tâche spécifique, c'est bien la volonté de ne pas laisser la linguistique en dehors des développements et des travaux sur le langage qui reste la première motivation.

8.3 Replacer la linguistique dans les approches : l'exemple de l'attribution d'auteur

Je vais tâcher ici de montrer par un exemple comment nous avons tenté un rapprochement avec une partie de la communauté du TAL pour laquelle les méthodes par apprentissage sont depuis très longtemps installées.

Début 2011, nous avons décidé au sein de l'axe TAL de CLLE-ERSS de participer à une des nombreuses compétitions qui structurent certaines parties de la communauté du TAL, dans le sillage des conférences MUC (pour l'extraction d'information) et TREC (pour la recherche d'information) qui furent les premières (voir frise chronologique en figure 8.1). Ces compétitions concernent des tâches clairement identifiées (recherche d'information, questions-réponses, fouille de texte, etc.) et leurs organisateurs proposent ainsi un banc d'essai commun pour permettre la comparaison des méthodes utilisées par les participants. Pour ma part, je ne connaissais que grossièrement les principes de ces événements, pour avoir participé

au second plan à une des tâches de la campagne CLEF pour la recherche d'information monolingue avec mes collègues de l'IRIT. La décision de participer à une compétition de ce type était initialement motivée par notre volonté d'« accroître la visibilité de l'équipe à l'échelle internationale » et de mettre à l'épreuve l'ensemble des compétences et des ressources disponibles au sein de l'axe TAL. Notre choix s'est porté, un peu au hasard du calendrier, sur la tâche d'attribution d'auteur, une de celles qui sont proposées chaque année depuis 2007 dans la cadre de l'atelier PAN (*Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*[6]). Le travail présenté ici a été réalisé en collaboration avec Basilio Calderone, Nabil Hathout, Franck Sajous et Assaf Urieli, mais a également impliqué les réflexions de tous les membres de l'axe TAL de CLLE-ERSS et bénéficié de leurs suggestions et intuitions (Tanguy *et al.*, 2011b).

8.3.1 Quelques mots sur la tâche et la communauté

Le principe d'attribution d'auteur sur la base de traitements automatiques a connu en France des épisodes notables sur des cas célèbres, notamment avec les travaux de Dominique et Cyril Labbé qui ont argumenté par des méthodes quantitatives (basées sur les fréquences lexicales) pour l'attribution à Corneille de certaines pièces signées de Molière (Labbé et Labbé, 2001) et entraîné de houleux débats dans la communauté tant littéraire que de l'analyse statistique des données textuelles. En dehors de questions de ce type, il semblerait toutefois que la tâche d'attribution d'auteur ait peu motivé la communauté française du TAL. Il s'agit pourtant d'une application intéressante sur plusieurs points :

- elle correspond à de vrais questionnements intéressant différents pans de la société, que ce soit comme on l'a vu le monde de la littérature, mais aussi la justice (avec l'identification des auteurs de courriers anonymes et le repérage des prédateurs sexuels sur les sites de discussion en ligne[7] par exemple) ;
- en tant que tâche pour l'apprentissage supervisé, elle correspond au cas typique de la classification d'un texte. Qui plus est, la phase d'annotation des données pour fournir un corpus d'entraînement ou de test est triviale, puisqu'il suffit d'indiquer pour chaque texte qui en est l'auteur (une information objective généralement facile à trouver) ;
- l'ampleur de la question posée par les traces objectives et quantifiables qu'un auteur laisse dans un texte s'étend sur tous les niveaux de l'analyse linguistique, du lexique et de la syntaxe employés à la structuration globale du texte.

La tâche d'attribution d'auteur de PAN pour l'année 2011 se basait sur un corpus de courriers électroniques extraits du corpus ENRON (Klimt et Yang, 2004). Il s'agit donc de données réelles[8], en anglais, aussi variées et bruitées (écriture condensée, citation d'autres messages, présence de données non textuelles, etc.) que ce que l'on peut trouver en ouvrant n'importe quelle boîte à lettres électronique.

Comme toutes les tâches de ce type, la compétition PAN se découpe en différentes phases :

1. publication des données d'entraînement (donc des courriers avec indication de leur auteur) ;
2. publication des données de test (non catégorisées) ;

6. Le site de l'atelier est : `http://pan.webis.de`
7. Ce type de tâche hautement utile est d'ailleurs au programme de PAN 2012.
8. Ce n'est malheureusement pas toujours le cas. Pour preuve, la majorité des données utilisées au sein de ce même atelier pour l'identification des plagiats provenait de perturbations automatiques de textes.

3. soumission des résultats obtenus par les participants sur les données de test ;
4. calcul des scores obtenus et publication du classement ;
5. conférence et présentation des détails des méthodes utilisées par les participants.

Comme je l'ai dit, le calendrier resserré de ces compétitions décourage de fait toute approche autre que celle recourant à l'apprentissage automatique.

Plusieurs sous-tâches étaient proposées, dont les variations étaient au final assez minimes (nombre de messages à traiter, nombre d'auteurs à considérer, présence ou non d'auteurs absents du corpus d'apprentissage).

Il existe une importante littérature sur les méthodes concrètes d'attribution d'un texte à un auteur en se basant sur des collections de référence dont les auteurs sont connus (voir Juola (2006) pour un panorama). Si les méthodes par apprentissage y sont systématiquement utilisées et très variées dans leurs utilisations, on peut remarquer que les descripteurs utilisés sont généralement assez pauvres : sur l'ensemble des publications des participants aux compétitions PAN des 5 dernières années, moins d'un tiers par exemple font état de l'utilisation d'un étiqueteur, et 10% d'un analyseur syntaxique. Par contre, les fréquences des unités lexicales et/ou des trigrammes de caractères sont pratiquement systématiquement employées [9]. Un inventaire fait par Koppel *et al.* (2009) pour les années antérieures confirme cette évolution, et donne en plus un aperçu de l'évolution des techniques d'apprentissage, notamment de la montée en puissance des SVM.

Nous avons donc voulu aborder la question par le biais de l'utilisation de traits linguistiques riches, afin notamment de mettre en œuvre les différentes techniques et ressources que nous avons l'habitude d'utiliser au sein de l'équipe, et d'en mesurer concrètement l'impact sur ce type de tâche (en espérant bien entendu qu'elle apporteraient un gain quantifiable par rapport aux méthodes basées sur des traits plus rudimentaires).

8.3.2 Traits linguistiques riches

Notre approche a donc consisté à définir et implémenter le calcul d'un grand nombre de descripteurs, variés tant par les unités auxquelles ils se rapportent que par les ressources ou méthodes de calculs qu'ils nécessitent. Si certains de ces traits proviennent d'intuitions sur la pertinence et le pouvoir discriminant des phénomènes visés, d'autres sont issus par contre d'un examen plus précis des données d'entraînement, en faisant notamment appel à de l'observation outillée des courriers eux-mêmes.

L'ensemble du corpus a été tout d'abord nettoyé puis étiqueté et analysé syntaxiquement en utilisant la chaîne de traitement Stanford CoreNLP (Klein et Manning, 2003).

Pour présenter la totalité des traits implémentés, nous pouvons les regrouper par niveaux :

- au niveau sublexical : la fréquence des trigrammes de caractères, des suffixes, des signes de ponctuation, des *smileys* ;
- au niveau lexical : la fréquence des formes lexicales, l'utilisation de majuscules, le taux de mots morphologiquement construits, la longueur des mots, la répartition des parties du discours, les erreurs d'orthographes, les contractions, les variations entre anglais britannique et américain, le degré de spécificité et le taux d'ambiguïté moyen des mots dans WordNet, la fréquence des entités nommées ;
- au niveau phrastique : la fréquence des bi- et trigrammes de parties du discours, la complexité syntaxique, la fréquence des relations syntaxiques de dépendance ;

9. Je remercie Basilio Calderone pour cet inventaire.

– au niveau du message : la longueur totale (en mots et en caractères), la fréquence des retours à la ligne et des lignes vides, les formules d'introduction et de conclusion, la cohérence sémantique du message en utilisant des données issues d'une analyse distributionnelle.

Le détail de ces traits est présenté dans Tanguy *et al.* (2011b), mais on retiendra à ce stade la grande variété, et également la redondance de ceux-ci.

La dénomination de trait *riche* s'applique dans notre terminologie aux seuls descripteurs dont le calcul fait appel à une ressource externe, ou à la projection de connaissances plus complexes qu'un simple découpage. Ainsi, les traits pauvres que nous avons utilisés regroupent les fréquences des trigrammes de caractères, de formes lexicales, de la ponctuation, et les indications de taille. Certaines des ressources utilisées (en plus de l'étiquetage) relèvent de données génériques sur l'anglais (lexique morphologique CELEX, réseau sémantique WordNet, tables de voisinage distributionnel Distributional Memory de Baroni et Lenci (2010)) mais aussi de ressources développées de façon *ad hoc* à partir des données d'entraînement (formules d'introduction et de conclusion), ou de sources dispersées sur le Web (contractions, américanismes). Si du point de vue des possibilités du TAL et de la linguistique, ces descripteurs sont au final assez rustiques, dans le cadre de ce type de tâche ils apparaissent comme sophistiqués.

Sans appliquer de procédure de filtrage en amont, nous nous sommes appuyés aveuglément sur le classifieur par entropie maximale (cf. 7.1.1.2, page 177) pour exploiter la masse d'information contenue dans l'ensemble de ces traits.

8.3.3 Résultats et questions soulevées

A notre grande surprise, nous avons été déclarés vainqueurs de la compétition, nos propositions ayant atteint les meilleurs scores pour les tâches principales (avec une f-mesure moyenne de 0,5). Un mauvais choix stratégique nous avait conduit à utiliser des arbres de décision pour une tâche corrélée (la vérification d'auteur) pour laquelle nous avons eu des résultats très décevants, mais nous avons vérifié a posteriori que l'utilisation du classifieur par entropie maximale nous aurait également placés en tête.

Il est toutefois difficile de tirer des conclusions constructives à partir de cette évaluation qui ne reste que quantitative. La raison principale de notre succès pourrait bien être la méthode d'apprentissage (qui semble peu utilisée, les grands classiques pour ce type de tâche semblent bien être les SVM), mais nous avons de bonnes raisons de penser que les traits linguistiques riches déployés ont joué un rôle crucial. Cela reste bien entendu à prouver, et c'est ce que nous sommes actuellement en train d'examiner de plus près, même si les modes d'observation d'un classifieur probabiliste sont, comme on l'a dit, assez peu maniables.

Le premier mode d'étude est celui qui consiste à mesurer le rôle individuel d'un descripteur par rapport à la tâche, en utilisant des indicateurs statistiques. Notamment, la notion de gain d'information consiste à mesurer la liaison entre le descripteur et la variable de classe, en utilisant un calcul d'entropie. C'est d'ailleurs la mesure qui est utilisée dans les arbres de décision pour décider du descripteur le plus discriminant à sélectionner pour un nœud. Par contre, cette mesure ne donne qu'une indication qui concerne un descripteur isolé, sans permettre de les évaluer efficacement en groupe. Or, même si les traits linguistiques « riches » que nous avons utilisés sont bien plus synthétiques que les fréquences lexicales ou de trigrammes, ils n'en forment pas moins des familles parfois assez nombreuses (type de fautes d'orthographes, de relations syntaxiques, etc.). De plus, aucun de ces traits ou familles de traits n'a au final la prétention d'être autonome pour ce type de tâche.

Le second moyen d'accès à une compréhension des mécanismes est d'effectuer des *tests de lésion*, autrement dit de répéter la même tâche, avec le même système sur les mêmes données, mais en supprimant des descripteurs et en mesurant la diminution de performance globale. Ces tests ont l'avantage d'être simples à réaliser (mais très gourmands en temps de calcul), et applicables à tout type de configuration (sous-familles de descripteurs). C'est ce type de travail que nous avons commencé à mener et qui semble confirmer les points suivants :

- le « gros » du travail de classification semble bien être effectué par les descripteurs de base, essentiellement les trigrammes de caractères : ce sont eux qui, isolément, en tant que sous-famille de descripteurs obtiennent les meilleures performances ;
- les traits riches apportent effectivement un gain significatif sur cette base. Le gain est sans doute disproportionné par rapport à l'énergie dépensée pour le calcul de ces descripteurs sophistiqués (sans compter celle nécessaire à la création des ressources que nécessite leur calcul), mais c'est *a priori* lui qui nous a permis de prendre la tête de la compétition ;
- en faisant varier la taille du corpus d'apprentissage, on se rend compte que les traits linguistiquement riches prennent l'ascendant sur les trigrammes lorsque les données sont moins volumineuses, comme le montre la figure 8.2 dans laquelle sont présentés les scores de précision globale obtenus sur une sous-partie des données (limitée à deux auteurs) pour différentes configurations de descripteurs. Cette différence d'efficacité s'explique logiquement par la nécessité pour les descripteurs pauvres d'atteindre une masse critique pour être efficaces. En revanche, la projection de connaissance linguistique est tout à fait apte à fonctionner à faible volume, puisqu'on peut les voir (naïvement ou statistiquement) comme une cristallisation de la connaissance accumulée par l'observation de grandes quantités de données (par les linguistes).

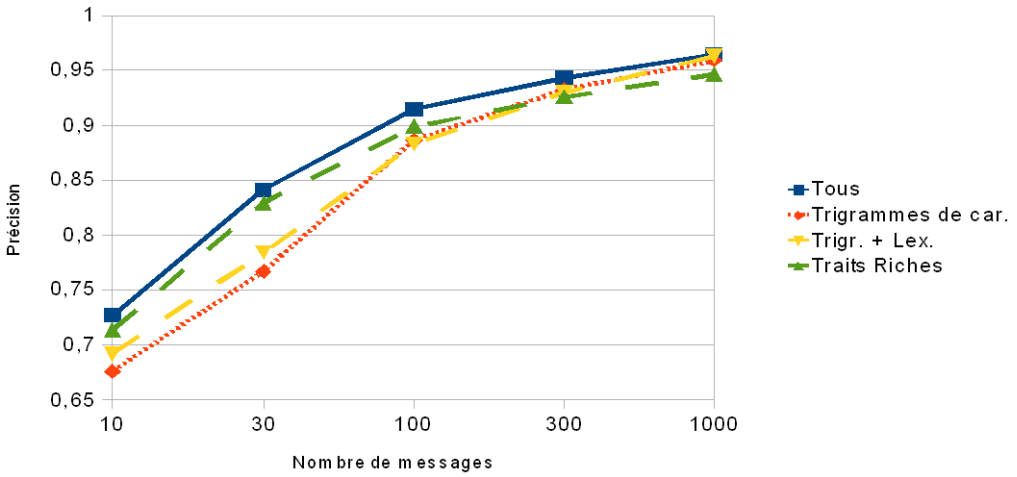

FIGURE 8.2 – Variation de la précision obtenue pour l'attribution d'auteur en fonction des types de descripteurs et de la taille du corpus d'entraînement

Quoiqu'il en soit, il reste encore des efforts à faire pour convaincre la communauté de l'utilité des ressources et méthodes linguistiques complexes pour ce type de tâche. Il est nécessaire pour ce faire de jouer le jeu de ce genre de compétitions, et d'apporter la linguistique

là où nous estimons qu'elle est utile, puisqu'il semble qu'on ne viendra pas la chercher.

8.4 Travail linguistique avec les méthodes par apprentissage

Ces quelques réflexions, ainsi que ces expériences avec les méthodes classiques de l'apprentissage et de la fouille de données m'ont permis de soulever un ensemble de questions relatives à la place de ces techniques dans l'exploration et l'exploitation des données langagières annotées. Plus précisément, je souhaite éclairer certains points concernant la place que ces méthodes prennent, à mon avis, dans la boîte à outils du linguiste outillé.

8.4.1 Comparaison avec les méthodes d'analyse statistique

Le premier point concerne l'apparente proximité entre les méthodes de fouille et celles proposées par la statistique descriptive que j'ai esquissée au chapitre 6. Comme on l'a vu dans les quelques exemples présentés en section 7.2 (page 182), la plupart des conclusions accessibles par des méthodes de fouille de données ont rejoint celles qu'une étude des corrélations entre les variables avait permis d'identifier.

En effet, il existe un parallèle direct entre la mesure d'une dépendance statistique et le fait qu'une variable (descriptive) va être discriminante pour la prédiction d'une variable-cible. Des associations régulières entre certaines valeurs sont en effet recherchées spécifiquement par les techniques d'apprentissage, et ce sont donc logiquement les variables les plus liées que l'on retrouve comme décisives dans les systèmes à base de règles (i.e. proches de la racine pour un arbre de décision, ou présentes dans les premières règles d'une liste).

Par contre, il n'existe pas d'équivalent direct avec les seuils de validation des tests statistiques d'hypothèse. La seule forme de validation proposée par les systèmes de fouille de données sont les indications relatives de l'efficacité d'une règle et, dans certains cas, l'efficacité globale du modèle. Cette dernière relève exclusivement des techniques d'apprentissage supervisé, qui peuvent être évaluées en termes de précision lorsqu'on les projette sur des données similaires à celles utilisées pour la phase d'apprentissage (généralement obtenues en sélectionnant une part aléatoire des données annotées pour pouvoir y comparer la décision du système sans qu'elles aient été utilisées pour construire le modèle).

On serait donc tenté de voir dans les techniques de fouille de données une version simplifiée des analyses statistiques. Leur popularité est sans nul doute due en grande partie à la simplicité de leur mise en œuvre, elle-même redevable aux efforts (parfois commerciaux) de diffusion au travers d'outils conviviaux. De plus, leur paramétrage est souvent lui aussi relativement simple, et peu de précautions d'emploi gouvernent l'utilisation d'une technique particulière, à la différence des tests statistiques (comme les exigences de distribution des variables).

Les statisticiens comme Friedman (1997) regardent la fouille de données comme une discipline totalement extérieure aux statistiques, apparue dans le meilleur des cas sous la poussée de l'informatique et de l'explosion du volume des données, et dans le pire sous celle du marketing. Le principal reproche de Friedman est l'abandon des techniques plus sophistiquées des statistiques au profit de méthodes robustes et rudimentaires appliquées à des données volumineuses. Ses craintes concernent même l'avenir de sa discipline, et le dévoiement des filières de formation et des étudiants (attirés semble-t-il par des postes mieux rémunérés dans les grandes corporations).

On serait également tenté de faire ici un parallèle avec le rapport que les linguistes entretiennent avec les méthodes quantitatives. Friedman a, par contre, la sagesse de considérer que

chaque camp doit faire un ensemble de concessions : s'approprier les avancées technologiques de l'informatique et les approches volumineuses qu'elles permettent pour les uns, et accepter de travailler plus finement sur des données moins volumineuses pour les autres.

A ce stade, le choix auquel est confrontée la linguistique concernant son outillage quantitatif pourrait être la suivante : doit-on privilégier des approches mathématiquement rigoureuses, mais nécessitant une connaissance plus approfondie des principes sous-jacents, et limitée à des quantités raisonnables, ou bien des techniques plus rudimentaires, conçues pour fonctionner sur de très grands volumes ?

Dans une situation idéale, on utiliserait effectivement les secondes en première intention, dans une phase de dégrossissage d'une exploration, avant d'affiner en focalisant sur des points plus particuliers avec des outils plus minutieux, afin de rejeter ou valider des tendances repérées dans la phase initiale.

Mais étant donnée la distance qui existe et semble se creuser entre la communauté des linguistes (même outillés, et habitués aux quantités de données et d'information), il semble donc qu'un choix doive être fait, et que celui-ci se porte sur les méthodes par apprentissage. On peut en résumer brièvement les principaux avantages :

- leur facilité d'emploi, due en grande partie à leur popularité en TAL qui tend à les rendre plus accueillantes, mais également au travail de vulgarisation (certes critiquable) qui a su les mettre à portée (et les vendre) à des secteurs d'activité variés ;
- leur capacité à gérer la masse de données, et le fait est que la linguistique est à la fois consommatrice et génératrice de grands volumes. On a vu notamment que ces quantités ne sont pas adaptées aux approches des statistiques classiques, notamment les tests d'hypothèses ;
- leur attractivité due au principe même de l'apprentissage automatique, héritière de l'aura qu'a su se donner l'intelligence artificielle malgré les promesses non tenues. Pour les étudiants en tout cas, le poids de ces aspects (avec le premier point) est énorme face à l'austérité et l'aspect effrayant des statistiques ;
- leur position privilégiée dans le TAL. Si les méthodes statistiques plus fines sont clairement mieux implantées dans certaines domaines de la linguistique (lexicométrie, psycholinguistique), on a vu que ce sont clairement les techniques de fouille et d'apprentissage qui occupent actuellement la scène.

Ces avantages ne doivent pas bien entendu cacher les grands problèmes de ces méthodes, ni nous faire rejeter les méthodes plus fondamentales des statistiques. Si un avis éclairé sur la rigueur mathématique des méthodes employées n'est pas à attendre de simples utilisateurs comme les linguistes outillés, on sait que les pièges sont nombreux et difficiles à détecter. Mais surtout, l'utilisation de méthodes de fouille et d'apprentissage entraîne également la production d'une masse d'information parfois difficile à interpréter. Si les méthodes probabilistes comme celles de l'entropie maximale ne permettent des observations que rudimentaires et peu satisfaisantes, les méthodes symboliques ont elles aussi une fâcheuse tendance à produire des quantités d'informations difficiles à digérer : que faire d'un arbre de décision doté de dizaines de nœuds, ou de règles d'associations produites par milliers ?

Les réponses viendront toujours des mêmes directions : la visualisation des données (point de vue global) et les tests spécifiques de la validité de certaines régularités (point de vue local), dans tous les cas avec un retour nécessaire aux données initiales. Pour ces deux cas, on a vu que les méthodes statistiques sont incontournables.

8.4.2 Multiplication des indices linguistiques

Plusieurs des travaux présentés dans ce chapitre et le précédent, qui ne se distinguent pas en ce sens d'autres expériences sur des phénomènes linguistiques complexes, se basent sur un très grand nombre de descripteurs. Cette multiplication est rendue possible par la facilité avec laquelle on peut calculer des indices de différents types et les projeter sur des corpus. Elle correspond à une approche expérimentale dans laquelle les intuitions linguistiques demandent à être confirmées empiriquement.

L'objectif de ce type d'approche peut être multiple : l'étude d'un phénomène particulier à partir d'observations en corpus (projet Annodis pour les structures du discours, thèse de Marion Laignelet pour les segments d'obsolescence), l'émergence d'une typologie d'unités (projet Rhecitas pour les contextes de citation, projet CAAS et thèse de Simon Leva pour les requêtes) ou la classification automatique (travaux avec CFH, attribution d'auteur).

Dans tous ces cas, qui se recouvrent d'ailleurs bien souvent, l'ordre des opérations est généralement le suivant :

1. Définir des indices calculables, et les projeter sur le corpus. La nature et les détails formels de ces indices peuvent provenir de connaissances linguistiques génériques en amont, ou d'une observation des données.
2. Calculer leur corrélation avec le phénomène visé, que ce soit isolément ou par faisceaux. Alternativement, les seules combinaisons d'indices peuvent être étudiées pour faire émerger une typologie.
3. Identifier les indices et les configurations pertinentes, les interpréter, les affiner.
4. Le cas échéant, construire sur cette base un dispositif permettant le repérage automatisé du phénomène ou la classification des unités.

Comme on l'a vu depuis les premières expérience de Biber (1988) pour étudier la variation, les techniques de TAL sont nécessaires pour systématiser et multiplier les indices à l'échelle d'un corpus entier. Les techniques quantitatives interviennent en aval pour faire le tri au sein de ceux-ci, si bien qu'il n'est nullement problématique de surcharger l'annotation, et d'arriver comme on l'a vu à des situations faisant intervenir des dizaines voire des centaines de descripteurs.

Bien entendu, cette multiplication des traits doit être adressée spécifiquement par les méthodes, si bien que certaines d'entre elles sont privilégiées. Il faut notamment qu'elles ne soient pas handicapées par la grande redondance entre les traits, et qu'elles soient capables de faire émerger des sous-ensembles stables de configurations en lien avec le phénomène visé.

Ce genre d'approche est surtout utilisé pour des phénomènes à l'échelle du texte pris dans son ensemble (classification de documents, en genre ou autre) ou à des phénomènes relevant du discours, comme l'a problématisé Péry-Woodley (2005), pour lesquels les indices sont de toutes façons multiples et insuffisants par eux-mêmes.

8.4.3 Place de la linguistique par rapport aux descripteurs

Au final, on voit donc se dégager une tendance qui permet d'éclairer la place des approches linguistiques dans l'utilisation de ce genre de techniques. La multiplication des traits pour la recherche d'indices stables ou pour des applications est en effet un indicateur du travail fait et restant à faire.

De ce fait, un certain éloignement de la méthode elle-même est compréhensible, et inévitable de par la distance culturelle avec les mécanismes mathématiques impliqués. Devant l'important éventail de techniques possibles, et de celles qui ne manquent pas d'être mises au point régulièrement, le choix est en fait assez peu important par rapport à celui de descripteurs adaptés et éclairants. Il reste donc, pour les acteurs du versant linguistique, à accepter le minimum d'obscurité qu'imposent les techniques par apprentissage, et à en tirer le meilleur parti, sans bien entendu les rejeter. Parallèlement, le choix d'une technique particulière et une familiarisation avec le fonctionnement de celle-ci est un atout important pour les premiers pas, quitte bien entendu à se tourner vers une collaboration avec des spécialistes de l'apprentissage.

Ce positionnement de la linguistique outillée est à mon avis d'autant plus nécessaire que, comme on l'a vu, les versants applicatifs (au moins eux) ont tendance à se détourner de l'utilisation de traits linguistiques riches. Pour reprendre le cas de l'attribution d'auteur, il semblerait en fait que la tendance croissante aille vers un appauvrissement des descripteurs. Pour preuve, dans les années 90 on trouve plusieurs travaux sur cette tâche qui utilisaient des données très riches issues d'analyses syntaxiques, comme par exemple ceux de Baayen *et al.* (1996) qui utilisent des règles de réécriture issues d'une analyse profonde. Il est en fait fort probable que l'évolution des techniques d'apprentissage et leur capacité accrue à digérer de très grands nombres de descripteurs soit une des raisons de l'abandon progressif de mesures synthétiques et qui condensent les informations linguistiques acquises par des analyses complexes, au profit de descripteurs très pauvres mais très nombreux, comme les trigrammes de caractères.

8.4.4 Se décomplexer par rapport aux méthodes

Je terminerai ce chapitre par l'évocation d'un signe à mon avis significatif de l'évolution des rapports des linguistes aux outils par apprentissage au sein du TAL : le détournement des méthodes.

Dans leur étude de la morphologie du Dholuo (une langue en danger parlée au Kenya), Pauw et Wagacha (2007) ont cherché à faire émerger des classes lexicales sur la seule base de leur ressemblance orthographique, en identifiant des schémas sous-jacents à ces classes. Le détournement effectué dans ce travail est d'utiliser un classifieur par entropie maximale comme une méthode non-supervisée, en utilisant le truchement suivant : chaque mot du lexique est décomposé en n-grammes de caractères de différentes tailles, ces n-grammes constituant alors les descripteurs. La classe visée pour l'apprentissage n'est autre que la liste des mots du lexique, chacun n'étant présenté qu'une seule fois : il s'agit donc bien d'une utilisation très atypique d'une méthode d'apprentissage supervisée, puisque chaque classe ne sera représentée que par un seul item. La phase d'apprentissage va cependant permettre de calculer, comme on l'a vu, un ensemble de poids relatifs à la répartition de ces n-grammes dans l'ensemble du lexique. Une fois le modèle construit, chaque mot est à nouveau présenté en entrée du classifieur. Bien entendu, si la prédiction va proposer le mot lui-même comme sortie du système, l'information pertinente dans leur approche sera la liste des autres mots du lexique, pondérés par des probabilités. En appliquant un seuil sur ces probabilités, il leur est donc possible d'extraire au final un réseau de relations de proximité morphologique entre les formes du lexique, et d'identifier ainsi des familles de mots. D'après leur expérience, la pertinence des classes ainsi construites est supérieure à ce que fournirait une mesure explicite de similarité comme la distance d'édition.

D'un point de vue pratique, il est clair que la méthode par entropie maximale est simple d'utilisation, et que l'opacité du système n'est pas un problème, puisque seules les sorties (pondérées par des probabilités) sont exploitées. Ce qui rend cette expérience possible est la robustesse de la méthode par entropie maximale à la redondance des descripteurs (puisque chaque séquence de lettres y est présentée plusieurs fois, avec la variation de la taille des n-grammes). La qualité des résultats est, elle, explicable par la prise en compte par le classifieur de la répartition et de la cooccurrence de n-grammes à l'échelle du lexique, puisque le modèle est construit sur l'ensemble des mots de celui-ci.

Le fait que de nouvelles utilisations des méthodes par apprentissage émergent dans la communauté est à mon avis un très bon signe, en ce sens que les techniques sont suffisamment mûres et intégrées pour qu'une certaine agilité (voire déviance) soit permise dans leur manipulation.

Il est donc possible qu'à terme les techniques par apprentissage soient utilisées par les linguistes comme des outils de base, au même titre que les étiqueteurs automatiques de textes, dont l'utilité est démontrée pour un très grands nombre de travaux en linguistique de corpus. La différence est pourtant très importante au regard de la situation actuelle : l'intégration de ces étiqueteurs a nécessité un long processus d'accommodation permettant notamment aux linguistes de bien percevoir leurs limites et leurs conditions d'emploi, et le cas échéant d'intervenir pour les adapter à leurs besoins. Cette accommodation requiert un minimum de familiarité avec leur fonctionnement, ne serait-ce que pour en maîtriser les biais. C'est cette familiarisation avec les méthodes par apprentissage qui reste à construire.

Conclusion

J'espère avoir pu donner dans ce mémoire un aperçu des points qui, à mon avis, doivent mériter l'attention et les efforts d'une partie de la communauté de la linguistique et du TAL afin d'aborder les rapides évolutions qu'elle est en train de connaître, et dont les conséquences sont déjà très sensibles. Je vais préciser dans cette partie conclusive trois points différents.

Je commencerai par un aspect important de mon métier d'enseignant-chercheur, en discutant plus spécifiquement des implications pédagogiques des différents sujets abordés, afin notamment de préciser la nature des compétences techniques et informatiques nécessaires et la place qu'elles doivent occuper dans la formation des étudiants en sciences du langage.

J'énoncerai ensuite quelques grands principes que j'ai pu dégager concernant le travail outillé sur les données langagières, qui ont surtout pour but d'éviter les principaux problèmes que j'ai pu identifier, et donner quelques lignes de conduite.

Je terminerai ce mémoire par une présentation des grandes lignes d'action que je compte personnellement suivre et développer, dans la continuité de mes travaux passés, mais dans des directions que ce nécessaire regard en arrière a permis d'identifier.

1 Implications pédagogiques des positionnements en recherche

Je n'ai que très peu parlé dans ce mémoire de mes activités d'enseignant, et je souhaitais donc leur rendre la place qu'elles méritent dans la présentation de mon travail.

J'ai milité (et pratiqué) au long de ma carrière d'enseignant pour l'accès des étudiants de sciences du langage aux techniques informatiques. Je vais en préciser ici les grandes lignes et les motivations en dégageant les principaux points que j'estime être importants, avant de soulever quelques questions concernant l'évolution de cette pratique.

1.1 Liste des compétences à acquérir pour les étudiants

J'ai tenté de dresser la liste des compétences techniques indispensables pour un travail sur les données langagières et qui doivent donc faire partie du bagage qu'acquiert un étudiant lors d'une formation en TAL. Dans la table 8.1, je les indique par catégories de compétences, en précisant en première colonne l'année de formation au cours de laquelle j'estime que ces compétences doivent être acquises. Notre expérience au département des sciences du langage concerne un début de formation en dernière année de licence (L3).

Il va de soi que ces compétences s'accompagnent d'une acquisition graduelle de la connaissance des techniques du TAL, des ressources disponibles et envisageables, et des enjeux scientifiques et technologiques.

Manipulation du matériau langagier	
1	*prévenir, diagnostiquer et résoudre les problèmes de codage* des fichiers de textes (codage des caractères et des fins de ligne)
1	*identifier et savoir convertir les différents formats de documents* pour au minimum en extraire le contenu textuel (et si possible, quelques méta-données et/ou informations structurelles)
Utilisation de l'outillage	
1	*appliquer des annotateurs génériques* (étiqueteurs, parseurs) en tenant compte du paramétrage de ceux-ci et des spécificités formelles des textes cibles
1	*convertir les sorties* d'un analyseur aux besoins d'un outil qui les exploite (concordancier, outil d'interrogation, extracteur spécialisé, tableur ou outil d'analyse statistique, base de données, etc.)
2	*projeter des ressources* sur un corpus (lexiques, marqueurs, patrons morphosyntaxiques, etc.)
2	*assembler des unités de traitement* en une chaîne complète, cumuler les informations notamment en utilisant un format XML
Exploitation	
1	*effectuer des recherches* et des extractions simples dans des fichiers de texte nu (segmenter, extraire des séquences de formes, calculer les fréquences)
1	*exploiter des annotations* automatiques pour rechercher et extraire des unités plus complexes
2	*évaluer un traitement*, en utilisant des mesures adaptées à la tâche, et en se basant sur un étalon ou sur des compétences de linguiste
2	*diagnostiquer* des fonctionnements erronés d'un système, et identifier un contournement ou une solution
Conception	
2	*déployer une méthode* de traitement raisonnablement complexe à partir d'une description du type de celles que l'on trouve présentées dans un article scientifique
3	*proposer une méthode* de traitement adaptée à un besoin nouveau, en se basant sur des solutions connues, mais nécessitant une articulation spécifique
3	*savoir dialoguer avec un informaticien* développeur, dans un travail d'équipe, en participant aux prises de décision nécessaires au développement d'un prototype

TABLE 8.1 – Liste des compétences à viser dans une formation de TAL en Sciences du Langage, par type et par niveau

Si certaines de ces compétences peuvent s'acquérir par le biais d'outils simples comme des éditeurs de texte ou ceux qui sont disponibles en ligne de commande (notamment sur les systèmes Unix), la maîtrise d'un langage de programmation ou de script est à mon avis la meilleure façon de les réunir en focalisant les apprentissages autour de la logique inhérente à ce langage.

1.2 Place de la programmation

Si l'apprentissage de la programmation fait partie de la grande majorité des cursus de TAL en sciences du langage, notre décision commune dans l'équipe pédagogique du TAL à l'UTM (composée de Marie-Paule Péry-Woodley, Cécile Fabre et moi-même) a toujours été de le renforcer au maximum. Dans les cursus où cette place est moindre, l'accent est généralement mis sur l'apprentissage des différentes fonctionnalités d'une plate-forme spécifique, comme

Nooj/Intex, peut-être GATE ou Linguastream. Sans vouloir dénigrer ces outils fort utiles, puisqu'ils permettent justement d'atteindre une opérationnalité tout à fait acceptable pour un ensemble d'objectifs d'exploitation des données langagières, je pense (comme je l'ai déjà dit) qu'ils entraînent également une distance dommageable avec le matériau de base, et qu'ils finissent de toute façon par rencontrer une limite dans les fonctionnalités envisagées.

C'est dans une certaine mesure ce qu'exprimait Biber au lecteur de son premier ouvrage de référence :

> *However, many interesting research questions involve investigating complex grammatical constructions or complex association patterns. Concordancing programs are not made for these sorts of investigations. For example, it is impossible with a concordancing program to conduct a thorough investigation of when* that *is omitted from* that*-clauses; and it would be even more difficult to look at the complex co-occurrences patterns of linguistic features in different registers. Instead, these investigations require computer porgramming skills [...] We encourage readers interested in pursuing corpus-based research on their own to take additional courses in computer programming.*
> (Biber, 1988)

Certes, 1988 se situe bien avant l'apparition des plate-formes génériques citées plus haut, mais je pense que si les limites des outils prêts à l'emploi ont largement reculées, les techniques du TAL et les modes d'investigation que l'on peut envisager ont eux aussi progressé.

Il est clair toutefois que l'enseignement de la programmation à des étudiants de sciences du langage nécessite un ensemble d'ajustements par rapport à ce qui est proposé aux étudiants des filières scientifiques. En plus des spécificités du matériau visé (en gros, les chaînes de caractères et les fichiers de texte), le bagage initial des étudiants est très différent, les volumes horaires disponibles bien plus faibles et la période de progression bien plus courte entraînent un sérieux resserrement. J'ai donc rapidement pris le parti d'un ensemble de raccourcis dans la présentation des principaux concepts fondamentaux de l'algorithmique et de la programmation. La principale conséquence est aisément visible dans le style de programmation, et le manque cruel d'élégance des programmes écrits par les étudiants (les réactions des rares « vrais » informaticiens ayant eu l'occasion de feuilleter *Perl pour les linguistes* en sont une preuve).

À ma décharge, je peux dire toutefois que le niveau d'opérationnalité atteint par les étudiants ayant suivi ces enseignements correspond bien à mes attentes, et certains les ont largement dépassées. Je peux également me réfugier dans les caractéristiques spécifiques du langage Perl, que nous avons choisi parce qu'il correspondait au point commun (à l'époque au moins) entre tous les collègues enseignants et chercheurs de mon entourage concernés par ces questions. Fortement critiqué sur ces aspects, Perl est un langage peu lisible, très peu rigoureux, et qui tend à pousser à l'excès les programmeurs au développement de programmes impossibles à maintenir sur le moyen terme. Toutefois, l'écriture de « moulinettes » rapidement obsolètes, répondant à un besoin ponctuel de triturage de données constitue l'objectif principal des approches du TAL que nous envisageons, comme l'indiquent les principales compétences listées ci-dessus. Malheureusement, il arrive aussi que de tels programmes se retrouvent toujours en activité bien longtemps après leur date de péremption, au grand dam de ceux qui doivent les utiliser ou pire, les modifier. Mais dans le cas normal, un programme écrit avec les compétences que j'ai tenté de transmettre n'a comme destin idéal que celui d'être une preuve de concept, cédant rapidement la place à un prototype écrit dans un langage plus

adapté, par un programmateur mieux armé.

Si la programmation constitue comme je l'ai dit le carrefour central des compétences techniques acquises, une d'entre elles est cependant à bien faire ressortir : la maîtrise des expressions régulières. Ce formalisme multi-usages de manipulation des chaînes de caractères en fait le couteau suisse du linguiste travaillant sur des données. J'ai eu, avec grand plaisir et à plusieurs reprises, l'occasion de rencontrer d'anciens étudiants ayant quitté la formation avant son terme, et qui m'ont avoué avoir gardé de mes enseignements ce seul aspect, mais avec un grand bénéfice quelle que soit leur activité ultérieure (pour peu qu'elle ait bien sûr un rapport avec la manipulation informatique).

1.3 Enseignement des techniques quantitatives

Comme j'ai eu l'occasion de le dire, le tournant des méthodes quantitatives doit urgemment être intégré dans la formation des étudiants, et c'est ce que nous sommes actuellement en train de faire à l'UTM.

Les techniques statistiques ont généralement mauvaise presse parmi les étudiants des filières les plus littéraires des SHS, et en linguistique en particulier. Que ce soit à cause d'un rejet plus général des choses mathématiques (appuyé sur les mauvais souvenirs d'étudiants généralement formés dans les filières littéraires du secondaire), de la volonté farouche de garder une âme à son travail et à son objet, et de ne pas céder aux pressions de la quantification, les obstacles sont multiples.

L'angle d'approche le plus adéquat semble bien, là encore, être du côté de la représentation visuelle. Les mémoires de master d'étudiants de sciences du langage (quel que soit leur domaine) ne manquent généralement pas de diagrammes, plus ou moins heureux et justifiés, même s'ils sont généralement exempts de calculs. Donner quelques règles simples pour visualiser les données est donc un objectif aisément accessible en première approche.

Pour la suite, des enseignements différenciés s'imposent : si tous les étudiants de linguistique ont des besoins en méthodes quantitatives, il y a de grande variation entre les domaines d'étude, de par les besoins, les objets et les habitudes de chaque partie de la communauté. Les tests statistiques par exemple, sont aussi incontournables pour les psycholinguistes que les méthodes par apprentissage pour les étudiants de TAL.

Concernant ces derniers plus précisément, il existe plusieurs voies possibles. La première consiste à exposer les grands principes et les familles d'outils et à donner un savoir-faire minimal sur les techniques les plus simples (techniques bayésiennes, extraction de règles), et quelques pistes pour envisager plus tard l'utilisation de techniques plus complexes (SVM, entropie maximale, CRF, etc.). L'autre consisterait à rentrer dans les détails d'une technique particulière, quitte à spécialiser les étudiants, mais en les rendant opérationnels face à une gamme raisonnablement étendue de situations. Actuellement mon choix s'est porté sur la première, préférant donc une phase de sensibilisation à une opérationnalisation trop restrictive, donc à l'inverse de ce que j'ai toujours prôné concernant les manipulations par la programmation. Mais l'objectif reste quand même de transmettre aux étudiants suffisamment de connaissances concernant leur environnement disciplinaire pour leur permettre d'y évoluer efficacement. De plus, l'enfermement dans une méthode particulière serait dangereux étant donné la rapidité d'évolution des techniques.

1.4 Des vertus pédagogiques de la technique et des projets

Je terminerai ces quelques réflexions sur mon rôle d'enseignant par des considérations plus générales, liées en partie avec ce que je pense être les aspects cachés des intérêts de l'enseignement plus technique et appliqué en sciences du langage sur lequel je me suis concentré.

Un des avantages, pour les étudiants, de disposer de compétences techniques opérationnelles est de leur permettre d'accomplir des réalisations allant jusqu'au bout d'une question posée. De la problématisation à l'évaluation des résultats obtenus, en passant bien entendu par la réalisation d'un programme dédié et son application à des données, j'ai toujours pris soin d'insister pour que les étudiants mènent une telle expérience à son terme, généralement dans le cadre de projets encadrés. Tout comme pour les stages, j'ai pu remarquer les bienfaits qu'apporte le fait d'avoir réalisé quelque chose, non seulement en terme d'apprentissage de savoir-faire, mais surtout d'accroissement de la confiance en leurs capacités.

Ce type de réalisation leur permet également d'avoir un point de vue bien plus précis sur le coût véritable de mise en place d'une application informatique, et leur fait relativiser les trop fréquentes formules du type « *avec cette ressource/méthode, on peut faire X* ».

Bien entendu, il convient de rester vigilant sur d'autres aspects, plus négatifs ceux-ci. Le premier danger est celui de la fascination que crée tout nouvel outil (surtout quand on le construit soi-même), et qui a une fâcheuse tendance à éloigner des objectifs initiaux, et pire, des données elles-mêmes. Il est donc systématiquement nécessaire de savoir recentrer l'observation sur les résultats, et non sur la technique qui a permis de les obtenir. De même, il est important pour les étudiants (comme pour nous) de ne pas se contenter des évaluations globales traduites par les indicateurs numériques comme le rappel et la précision, et savoir toujours revenir au matériau langagier brut, celui par lequel tout a commencé.

2 Quelques principes pour le travail outillé sur les données langagières

La panoplie de travaux présentés dans ce mémoire montre bien la diversité des questions qui sont soulevées par des données linguistiques de natures très différentes, et rendues encore plus disparates lorsqu'elles ont été la cible de processus informatisés pour leur annotation, leur exploration, leur représentation ou leur analyse.

L'outillage informatique entretient au final des rapports contrastés et quelques fois paradoxaux avec l'étude du langage. Les bienfaits apportés par la possibilité de manipuler des quantités massives ne sont plus à démontrer : le changement d'échelle a rendu visibles de nouveaux phénomènes et éclairé bien plus nettement ceux qui étaient déjà connus ; les méthodes d'observations ont permis de repérer de nouvelles caractéristiques des unités linguistiques et de relier des questionnements jusqu'ici étudiés séparément.

Par contre, les techniques et les outillages ont des effets négatifs sur les modes d'approche de ces mêmes données : les outils d'exploration imposent une distance parfois regrettable avec le matériau langagier, les annotations automatiques induisent des biais qui peuvent avoir des conséquences importantes sur les observations, les outillages nécessitent une compétence technique qui peut être rebutante pour les linguistes, et enfin les techniques quantitatives comme celles basées sur l'apprentissage et les statistiques ajoutent une certaine opacité dans les manipulations et les observations.

Quoiqu'il en soit, je pense pouvoir tirer ci-dessous quelques principes plus généraux que la

réunion rétrospective de ces cas particuliers m'a fait cerner. Les quelques principes généraux que j'évoque ci-dessous s'appliquent donc aux situations où, face à un nouveau matériau à étudier, ou pour aborder une nouvelle question linguistique, une approche outillée est envisagée, en interaction avec les connaissances préalables disponibles, et dans le cas général dans le cadre d'une collaboration avec un chercheur moins familier avec les outillages en question.

2.1 Appréhender les données directement

La première condition de ces approches est l'existence d'un rapport direct avec les données à représenter. Cela paraît évident, mais c'est parfois une opération complexe qui nécessite un examen attentif des caractéristiques de celles-ci (que rien d'autre que la lecture directe d'un échantillon ne peut apporter). Dans certains cas il est également nécessaire de discuter en détails avec le « fournisseur » des données, notamment pour en découvrir les caractéristiques externes qui peuvent expliquer certains phénomènes qui ressortent de la visualisation ou de l'analyse. Les cas les plus complexes sont bien entendu ceux où un processus d'annotation préalable a été mis en place : la complexité des structures du discours annotées dans Annodis en est un bon exemple, comme l'est celle des relations sémantiques issues de l'analyse distributionnelle projetées sur des textes.

La connaissance approfondie des données permet par la suite de trouver des exemples probants, de tester le comportement des processus d'annotation et de visualisation sur des configurations particulières, et aussi d'adapter les traitements à leurs spécificités.

2.2 Comprendre les besoins

À de très rares exceptions près, je n'ai jamais été confronté à une demande précise de visualisation ou d'analyse nécessitant une approche innovante (contrairement aux étapes d'exploration de corpus, pour lesquelles on a plus facilement une vision précise des énoncés recherchés). Lorsque le besoin est clairement identifié et décrit, c'est généralement qu'il s'agit de réappliquer (en l'aménageant) une méthode déjà éprouvée, ou d'utiliser un outil disponible. Pour tous les autres cas, il faut s'adapter, et faire des propositions spécifiques aux données et au problème dégagé. C'est sans aucun doute la partie la plus enrichissante d'un travail systématiquement collaboratif, puisqu'elle ne signifie rien de moins que de rentrer dans une problématique de recherche, et d'y prendre une part active, alors que la seule fourniture ou conduite d'un appareillage est généralement un peu frustrante.

Des phases de dialogue avec les collaborateurs doivent donc être ménagées aux différents moments de l'étude, et doivent s'appuyer sur un échange doublement enrichissant. Avant tout déploiement d'une technique, il est important de bien comprendre les modes de travail, les théories sous-jacentes, les enjeux et les nouveautés de l'étude. La difficulté principale est de savoir s'arrêter à temps dans la plongée dans un nouveau domaine d'étude, et de savoir en identifier les principales caractéristiques nécessaires à l'établissement d'une solution satisfaisante pour tous.

2.3 Prendre en compte les connaissances déjà acquises

Le troisième principe découle du précédent, mais il est important de le préciser : dans tout travail d'investigation sur des données il est nécessaire d'avoir une familiarité avec les connaissances préexistantes. Nombre de travaux de ce type, mais surtout l'utilisation de techniques de fouille de données ont souvent comme seul résultat de faire ressortir des informations déjà

connues. Même si parfois l'objectivation et la quantification de ces évidences est utile sur le plan argumentatif (et rassurante quant à la méthode utilisée), le but est évidemment de faire émerger de nouvelles connaissances. Ainsi, retrouver comme paramètre le plus discriminant pour un type particulier d'unité un trait qui est en fait définitoire ne présente que peu d'intérêt.

Une des phases critiques pour la représentation et l'analyse quantitative est en effet la sélection des paramètres à intégrer : cette phase est à mon avis un des points d'articulation les plus délicats entre la connaissance des données et celle des méthodes.

2.4 Ne pas s'éloigner des données

La plupart des approches d'analyse des données se situent après un ensemble d'opérations parfois très complexes qui ont comme fonction d'éloigner la représentation des données « brutes ». Il peut s'agir d'une annotation sophistiquée, d'une représentation graphique abstraite, d'une classification intermédiaire, d'un filtrage ou d'un applatissement des données pour les rendre gérables par une méthode, etc.

Dans l'idéal, il est préférable d'effectuer un éloignement progressif, et de contrôler cette progression par un retour toujours possible aux données initiales, comme l'observation d'un segment de texte particulier, ou l'examen des instances d'une classe lexicale. Ces retours permettent à la fois de contrôler les étapes de calcul et bien entendu d'interpréter les résultats observés par la méthode.

Si ces retours sont généralement bien pris en compte dans les méthodes de visualisation, il s'agit bien souvent d'un des points faibles des approches quantitatives.

2.5 Faire attention aux biais

Toute utilisation d'un outil d'annotation introduit un ensemble d'erreurs et d'approximations, tout calcul statistique masque des phénomènes épars, et toute représentation visuelle réduit par principe les dimensions des données sous-jacentes. Ne pas prendre en compte certains aspects peut avoir des conséquences négatives, surtout dans ce genre de situation où la nouveauté d'une représentation ou d'une analyse entraîne un certain enthousiasme à l'interprétation.

Si le danger de ces biais est généralement bien connu en ce qui concerne les procédures d'annotation automatique (en grande partie par le recul que permettent des années d'utilisation, comme c'est le cas pour les étiqueteurs morphosyntaxiques) et les analyses statistiques (grâce aux expériences accumulées par les techniques fondamentales), ceux des techniques par apprentissage sont plus difficiles à appréhender.

Les linguistes de corpus ont appris à accepter le bruit que produit toute exploration outillée des données (et donc à en dépouiller patiemment les résultats), et dans une moindre mesure à considérer que toute recherche automatisée entraîne également du silence (et à revenir ponctuellement à des modes d'interrogation plus directs). Les moyens de lutter contre ces deux types de problèmes sont plus difficiles à envisager lorsque l'opacité des techniques ne permet pas d'envisager une vérification des résultats intermédiaires, si bien que seule une meilleure compréhension des méthodes peut permettre d'y remédier.

Les représentations graphiques ont également la fâcheuse tendance à faire apparaître des artefacts de représentation, dont l'interprétation est inévitable (ce que Rastier appelle la « condamnation au sens »). Si certaines méthodes statistiques, comme les analyses factorielles,

disposent d'un ensemble de garde-fous (comme la prise en compte des indications pour chaque axe de la part de complexité qu'ils traduisent), lorsque de nouveaux modes de visualisation sont mis en place, le danger est bien plus grand [10]. Les meilleures protections sont alors la connaissance des données et des phénomènes étudiés, et la multiplication des points de vues et des méthodes d'observation.

2.6 S'approprier les outils

Ce dernier principe est une conséquence directe des précédents, et concerne plus spécifiquement l'outillage (au sens large) que toute approche sur des données langagières numériques rend à la fois nécessaire et bénéfique.

La première remarque concerne la nécessité, comme je l'ai déjà dit, de s'accaparer les techniques et les méthodes. Cette prise en main va bien entendu au-delà de la simple faculté technique à les « faire tourner » (même si les efforts pour ce faire sont parfois très importants). Elle concerne bien plus la prise en compte des principes sous-jacents, des limites dans leurs applications, des biais induits et des modes d'interprétation des résultats.

La multiplicité des besoins, qui varient d'une étude à l'autre, et la nécessité d'aborder une même question sous différents angles de vue encourage, de plus, la maîtrise de plusieurs techniques, puisqu'aucun outil, aussi sophistiqué soit-il, ne pourra aborder de façon satisfaisante l'ensemble des configurations. Le choix de développements (théoriques ou informatiques) *ad hoc* va donc de soi, et les approches intégrées ne peuvent guère traiter que des situations de très bas niveau concernant des problèmes déjà bien balisés, comme l'exploration des phénomènes lexicaux dans les textes, que de nombreux outils de plus en plus complexes en lexicométrie permettent d'aborder. Mais même dans ce champ-là, la philosophie actuelle telle qu'elle est par exemple exprimée par Serge Heiden dans le cadre du projet Textométrie (Heiden *et al.*, 2010) va vers la mise en place de modules articulés et pour certains interchangeables. De plus, les questions de fond sur le rôle de la visualisation y sont également soulevées : même si des principes généraux sur la cible et la nature de la représentation sont identifiables, leurs instanciations doivent toujours s'adapter à des situations sans cesse nouvelles.

3 Perspectives et plan d'action

Je termine ce mémoire en récapitulant une série de points que je considère comme prioritaires pour les prochaines années. Sans avoir la prétention de croire qu'ils sont centraux pour l'évolution du TAL et de sa collaboration avec les travaux de linguistique, je vais préciser le rôle que j'estime être le mien face à ces questions.

3.1 Remplir mon rôle de passeur pour les approches quantitatives

J'ai déjà eu l'occasion de me qualifier comme passeur, en tant qu'informaticien de formation évoluant désormais dans un milieu de linguistes. Cette fonction ne se limite bien entendu pas à « vendre » des techniques et des outils d'analyse ou de description, mais au contraire de rentrer dans une problématique afin d'identifier un rapport productif entre les besoins

10. À titre d'exemple, dans les représentations des transitions entre thèmes dans les consultations médicales, un des graphes que j'ai dessinés, dans la veine de celui de la figure 5.2 (page 113), faisait apparaître au premier plan une catégorie anecdotique (*vie personnelle du médecin*) à cause de la systématicité de ses enchaînements.

et les solutions envisageables (ceci dit, c'est peut-être ce que l'on demande justement à des commerciaux...).

Mon activité initiale s'est surtout concentrée sur les techniques informatiques fondamentales pour le TAL : manipulation des données langagières brutes et annotées, techniques d'exploration et d'interrogation, modalités de représentation, analyses quantitatives. Je me suis, autant que faire se peut, également donné comme règle d'apporter des compétences directement utilisables par les linguistes n'ayant pas forcément de formation technique, notamment par des approches pédagogiques comme la rédaction de l'ouvrage *Perl pour les linguistes* avec Nabil Hathout (Tanguy et Hathout, 2007).

L'étape suivante concerne à mon avis la familiarisation de la même communauté avec des approches quantitatives, qu'il s'agisse des méthodes statistiques traditionnelles ou des méthodes par apprentissage. On regrettera par exemple l'absence actuelle d'un ouvrage en français traitant de ces questions, comme le fait par exemple Gries (2009), qui prend appui sur de véritables exemples issus des préoccupations langagières et donne des solutions techniques par le biais d'un outil directement utilisable (R). Cette situation est d'autant plus étonnante que la plupart des autres disciplines des SHS disposent de telles références (même le sport, voir (Champely, 2004)).

La tâche est d'autant plus aisée à mon avis pour les méthodes par apprentissage que certaines de celles-ci (celles à base de règles notamment) sont au final assez simples à comprendre et à mettre en place, et peuvent aisément s'appliquer à des données déjà rassemblées. On peut donc espérer qu'elles deviennent à terme un des éléments du bagage méthodologique et technique des linguistes travaillant sur des données.

Mais ce rôle de passeur est malheureusement un rôle qu'il est difficile de tenir dans la durée. Si ma formation initiale d'informaticien m'a permis de le faire pendant plus d'une dizaine d'années, je suis bien obligé de reconnaître que l'évolution des techniques et des outils est difficile à suivre, du moins en plus de l'évolution des questions propres à la linguistique. Par exemple, le langage Perl est sans doute en train de perdre la place privilégiée qu'il a pu occuper jusqu'au début des années 2000 en tant qu'outil incontournable pour la manipulation des données langagières (Python est actuellement sur le devant de la scène, notamment avec des bibliothèques dédiées comme NLTK [11]). S'il est aisé, comme l'ont prouvé plusieurs des étudiants que j'ai pu former, d'apprendre un nouveau langage de programmation une fois qu'on en maîtrise à peu près un, les mécanismes et les réflexes à acquérir évoluent également.

Concernant les méthodes par apprentissage, ma formation initiale et ma connaissance des techniques modernes ne sont pas suffisantes pour jouer ce rôle au même niveau que pour la programmation et ma position sera donc nettement plus proche de celle d'un utilisateur curieux et peu farouche. De fait, mon environnement actuel (et sans doute futur) m'a sans doute fait glisser progressivement du producteur informaticien au consommateur linguiste, et le moment est sans doute proche où il me faudra passer... la main.

Se pose alors les questions plus larges des conditions d'émergence des nouveaux passeurs. La complexification des techniques et le cloisonnement qu'elles induisent dans les formations sont autant d'obstacles, tout comme l'étanchéité disciplinaire qui impacte les recrutements. Je comprends mieux, étant arrivé au stade de ma carrière où je dois intervenir dans ce type de décisions, l'audace de ceux (et je les remercie) qui ont osé me recruter, avec mon bagage d'ingénieur informaticien, dans un laboratoire et surtout un département de linguistique. Jusqu'à

11. Natural Language ToolKit, qui regroupe à la fois des données génériques, des outils d'annotation automatique et des modes d'exploitation. `http://www.nltk.org/`

preuve du contraire, je ne vois cependant pas d'autre solution que d'accepter l'entrée dans une communauté scientifique et pédagogique d'un individu issu d'un milieu différent : les collaborations interdisciplinaires (aussi riches soient-elles) comme les projets finissent généralement par devoir répondre à des exigences disciplinaires divergentes.

3.2 Jouer dans la cour du TAL moderne

La familiarisation des linguistes avec les techniques par apprentissage a, comme je l'ai dit au chapitre précédent, deux intérêts. Le premier est la possibilité que ces techniques offrent d'apporter un regard nouveau sur des données complexes, permettant de dégrossir en première intention et de mettre au jour des régularités. Le second est de permettre aux linguistes de (re)prendre pied dans le TAL tel qu'il est actuellement, et de permettre un dialogue plus constructif avec les acteurs du TAL statistique.

Ce dialogue nécessite pour les linguistes d'aller au-devant des préoccupations du TAL applicatif, comme ils ont déjà depuis longtemps su le faire, en apportant plusieurs aspects avantageux :

- la volonté et les compétences pour observer plus finement les données, là où cet aspect essentiel du travail sur le matériau langagier a malheureusement commencé à disparaître (au moins en apparence et en tant que méthodologie) des activités du TAL. Les capacités à analyser manuellement des données, même volumineuses, par le biais d'outils simples d'accès aux données sont une source de renouvellement de questions, et dans certains cas des moyens incontournable d'éviter les biais dans les approches purement quantitatives, et de proposer des solutions ;
- la connaissance des phénomènes langagiers pertinents dans ces travaux, en lien avec le point précédent. Cela se traduit concrètement par une meilleure compréhension des articulations entre les descripteurs des données et les informations visées. Plus concrètement, ces connaissances permettent d'orienter les choix de descripteurs, voire d'injecter des règles spécifiques si la technique utilisée le permet (comme dans le cas de l'apprentissage structuré) ;
- l'apport de nouveaux descripteurs, notamment ceux qui sont issus de travaux exploratoires en linguistique descriptive, qui peuvent être basés sur des ressources développée sur corpus, et dont on peut espérer qu'ils se montrent stimulants et éclairants sinon utiles.

3.3 Exploiter la distribution des phénomènes dans les textes

Je pense avoir montré la nécessité de déployer des techniques de visualisation pour aborder la complexité des données langagières, surtout lorsque celles-ci reçoivent des annotations qu'il est nécessaire de croiser pour en percevoir les mécanismes d'interaction. Si une grande partie des efforts nécessaires dans cette direction consiste à développer des outils parfois sophistiqués, notamment lorsqu'ils mettent en place des procédures d'interaction avec l'utilisateur (et cet aspect est absolument central), d'autres pistes peut-être moins exigeantes sont à explorer.

On a pu voir que les représentations visuelles permettent d'identifier des formes particulières qui, si elles ne sont pas totalement caractérisées à la suite de ces observations, n'en ont pas moins une existence objectivable. L'exemple le plus probant de ce type d'approche est sans aucun doute la segmentation thématique telle que l'a présentée Hearst (1997), et comme on peut en voir un exemple dans la figure 5.5 (page 120) : ce sont les points de la

courbe qui correspondent à des baisses notables de la cohésion lexicale qui sont repérés dans la méthode de *Text Tiling* comme étant des ruptures thématiques dans le texte. Je pense que de nombreux autres phénomènes peuvent être abordés de cette façon ; j'en ai montré quelques exemples au chapitre 5, par exemple pour repérer les différentes phases d'une consultation médicale (figure 5.11, page 129).

L'étape suivante consiste alors en une généralisation de ce principe : définir les configurations topologiques de tels schémas de disposition dans un texte et automatiser leur détection. Cette opération passe par des procédures de reconnaissance des formes, rejoignant ainsi les propositions de Rastier (1991) qui voit dans cette approche une alternative justifiée au calcul *bottom-up* dès lors que la sémantique est en jeu. Peu de travaux semblent s'être orientés vers ce type d'approche, les plus avancés me semblent être ceux qui proposent une extension de la méthode des rafales, développée pour l'analyse de contenu par Lafon (1981) et qui vise le repérage dans un texte des zones de répétition d'unités lexicales. Ce principe a notamment été implémenté dans le logiciel Tropes, et des travaux plus récents proposent de les réutiliser pour des unités plus complexes que les simples formes lexicales (voir par exemple la notion de *motif* dans (Longrée et Mellet, 2010), et les premiers principes de la topologie textuelle, voir Mellet et Barthélemy (2009)).

Une autre piste similaire consiste à faire émerger des classes de documents sur la base de la répartition d'un même phénomène, par exemple par le calcul d'une corrélation (d'un texte à l'autre) entre les courbes qui les représentent. Les applications que j'ai évoquées dans ce mémoire concernent autant la classification des articles scientifiques en fonction de la répartition des appels de citation (voir 5.2.3, page 126) que des consultations médicales en fonction du partage de la parole entre médecin et patient (voir 5.2.4, page 128) ou encore des thématiques ou types d'incidents en fonction de leur répartition chronologique (voir 7.3.2, page 191). Une des pistes que j'ai commencé à explorer s'appuie sur des travaux sur les séries temporelles courtes, telles qu'elles sont étudiées dans des domaines assez éloignés comme la finance et la biologie (Todorovski *et al.*, 2002).

Dans les deux cas bien entendu, ces mesures automatisées n'interviennent qu'après une phase d'analyse minutieuse des données qui nécessite des outils de visualisation et des interprétations en lien avec le texte lui-même. Les calculs permettant de rapprocher ou typer ces configurations spatialisées découleraient donc directement des observations globales que seule la visualisation permet, et ne feraient donc qu'automatiser (par des calculs locaux) leur repérage sur des données plus massives. Ce passage nécessaire à une définition calculable d'une forme définie visuellement est bien entendu un appauvrissement de la structure repérée, comme c'est habituellement le cas dans toute modélisation sémantique.

3.4 Mieux collaborer avec les autres disciplines

On l'a vu à plusieurs reprises, j'ai été confronté régulièrement à des demandes de collaboration issues d'autres champs que la linguistique, et plus particulièrement à des besoins appliqués issus du monde de l'entreprise ou des questions plus théoriques provenant d'autres sciences humaines.

Il est très satisfaisant de voir que la linguistique outillée et le TAL sont de plus en plus à même de répondre à ces questions, même si leur visibilité est à améliorer. La maturité des techniques d'appréhension des données, la faculté de s'adapter à différents types de documents, ou encore de pouvoir faire appel à des données issues de sources génériques (comme les corpus ciblés issus du Web) font que les réponses aux premiers questionnements qui nous

sont adressés sont généralement assez rapides pour susciter l'intérêt et la collaboration. La maîtrise des méthodes statistiques, qui constituent de fait une sorte de *lingua franca* entre les disciplines (du moins entre certaines), facilite également ce dialogue entre des pratiques par ailleurs différentes. Les outils de visualisation forment, eux aussi, un médium facilitateur, à la fois comme proposition de réponse face aux questionnements initiaux, mais aussi comme moyen de comprendre plus précisément les données, les connaissances déjà acquises sur la question, et d'affiner les objectifs.

J'ai déjà parlé (voir section 5.3, page 131) des efforts à faire, tant à mon niveau qu'à celui de la communauté du TAL, pour approfondir et développer ces techniques de visualisation. Je souhaite néanmoins insister sur le type de réflexion qu'implique la mise au point d'une représentation graphique particulière et la nécessité d'expérimenter différents choix visuels. Je tiens par exemple à préciser que les frises chronologiques que j'ai placées à différents endroits de ce mémoire ont été obtenues après de nombreuses tentatives et des interactions avec mes collègues. Les choix finaux, bien entendu discutables, induisent un point de vue particulier sur les événements représentés. Même pour un objet aussi classique, plusieurs des paramètres sont le lieu d'alternatives ouvertes : la disposition verticale des événements (qui doit répondre avant tout à un critère de lisibilité), le marquage des « tournants » disciplinaires dans la frise de la figure 0.1 (page 25), la surimpression des courbes, etc. Le résultat final, comme tout travail de présentation, est donc un compromis entre la faisabilité technique, la nature des informations à présenter et surtout les différentes possibilités d'interprétation qu'elles induisent. Ce sont des questions de ce type qu'aborde spécifiquement la communauté disciplinaire de la visualisation de l'information à travers une approche scientifique désormais institutionnalisée, et dont je pense qu'il est important de se rapprocher.

On voit donc que ces méthodes, ainsi que les techniques dont se sont récemment dotés le TAL et la linguistique outillée sont des atouts majeurs pour répondre à des questions interdisciplinaires. Ce sont par exemple ces techniques qui permettent d'intégrer aisément les données dont disposent généralement nos interlocuteurs, et de les enrichir par des analyses linguistiques. Un cas exemplaire est le tout nouveau projet RESOCIT [12], dirigé par Béatrice Milard du laboratoire LISST et auquel je participe. Ce projet se propose d'étudier les différentes dimensions des relations sociales entre chercheurs, notamment telles qu'elles s'expriment par la citation d'une référence bibliographique dans un article scientifique. Les différentes questions de recherche soulevées ici concernent notamment, pour la part relevant de mon intervention dans ce projet, à confronter les caractéristiques linguistiques des appels de citation, non plus avec des fonctions rhétoriques comme dans RHECITAS, mais bien avec les intentions véritables de l'auteur citant. Les raisons d'une citation vont en effet bien au-delà du simple document dans lequel elle s'insère : elles peuvent correspondre à la reconnaissance de collaborations passées, mais aussi à des invitations à des travaux communs, à des rivalités scientifiques, etc. (Milard, 2011). Ces informations ne peuvent être obtenues que par les modes d'investigation de terrain de la sociologie (entretiens individuels avec les auteurs autour de leur publication), et la typologie ainsi proposée peut ensuite être aisément confrontée aux apports de l'étude des contextes linguistiques.

On voit à travers cet exemple, mais dans tant d'autres également, que ce type de collaboration va bien au-delà d'une simple valorisation des apports de la linguistique aux autres disciplines. Ces dernières peuvent en effet à l'inverse apporter une forme de validation et

12. Citations scientifiques et réseaux sociaux : étude des dynamiques relationnelles impliquées dans la production et la diffusion des publications scientifiques.

un regard extérieur sur des objets linguistiques, qu'il serait difficile, voire impossible, d'aller chercher par nous-mêmes. Dans le cas des citations, on a vu que les travaux d'analyse des publications se heurtaient à la difficulté de valider les typologies induites : un projet comme RESOCIT vise justement à apporter des réponses qui permettront d'infléchir les modes d'investigations linguistiques autour de ces questions. Nombre de méthodes évoquées dans ce mémoire peuvent aisément être appliquées avec profit à ce genre de situation : l'analyse de données, mais surtout la visualisation de dimensions variées autour du document, sont autant de lieux propices à la confrontation de points de vue différents.

Liste des projets de recherche

RESOCIT : *Citations scientifiques et réseaux sociaux : étude des dynamiques relationnelles impliquées dans la production et la diffusion des publications scientifiques*

Période : 2012-15
Financement : Agence Nationale de la Recherche, programme Blanc (SHS 1)
Partenaires : CLLE (CNRS / Université de Toulouse 2), IRIT (CNRS / Université de Toulouse 3), LERASS (Université de Toulouse 3), LEREPS (Université de Toulouse 1), LISST (CNRS / Université de Toulouse 2)
Coordinateur : B. Milard (LISST)
Rôle : Responsable de la tâche d'analyse des contextes linguistiques des citations

Ce projet propose d'aborder, par le biais d'une méthode mixte (qualitative et quantitative) diverses questions autour de la notion de citation dans les publications scientifiques. Principalement ancré dans la sociologie des sciences, RESOCIT positionne la relation de citation dans le cadre plus vaste des relations entre les chercheurs d'une communauté. Les méthodes déployées vont aborder ces questions par le biais d'entretiens ciblés de chercheurs autour d'une de leurs publications, l'étude du processus de production scientifique et de la diffusion des articles, mais aussi à travers l'étude linguistique des contextes de citations (tâche dans laquelle je suis impliqué).

Voir : conclusion, section 3.4 (page 231)

CITRI : *Exploitation des citations dans les articles de SHS pour la recherche d'information*

Période : 2011-14
Financement : PRES Université de Toulouse & Région Midi-Pyrénées
Partenaires : CLLE (CNRS / Université de Toulouse 2), IRIT (CNRS / Université de Toulouse 3)
Coordinateur : L. Tanguy (CLLE) & J. Mothe (IRIT)
Rôle : Co-responsable avec J. Mothe, encadrant de la thèse de Simon Leva.

L'objectif est, par le biais d'une analyse linguistique utilisant des techniques de traitement automatique des langues, de mettre en place une méthode opérationnelle permettant d'exploiter les relations de citation entre des publications scientifiques dans le domaine des sciences humaines et sociales afin de faciliter leur exploitation et la recherche d'information par des usagers.

Voir : 7.3.3 (page 192)

Détection de signaux faibles dans des bases de données textuelles de rapports d'incidents

Période : 2011-14
Financement : Convention CIFRE
Partenaires : CLLE (CNRS / Université de Toulouse 2), Conseil en Facteurs Humains (Toulouse)
Coordinateur : M.-P. Péry-Woodley
Rôle : Co-responsable avec M.-P. Péry-Woodley, encadrant de la thèse de Nikola Tulechki.

L'objectif principal de cette thèse est de proposer des méthodes automatiques et semi-automatiques de détection de signes d'alerte précoces de risques émergents parmi les documents contenus dans de volumineuses bases de données textuelles de rapports d'accidents et incidents.

Voir : 7.3.2 (page 188)

CAAS : *Contextual Analysis and Adaptive Search*

Période : 2010-14
Financement : Agence Nationale de la Recherche, programme Contenus et Interactions
Partenaires : CLLE (CNRS / Université de Toulouse 2), IRIT (CNRS / Université de Toulouse 3), LIA (Université d'Avignon)
Coordinateur : J. Mothe (IRIT)
Rôle : Responsable pour CLLE

L'objectif de ce projet est de prendre en compte différents éléments contextuels dans un système de recherche d'information textuelle. Les dimensions du contexte sont : le besoin d'information (requête exprimée et caractéristiques de l'utilisateur), la collection documentaire et les paramètres internes du moteur de recherche.

Voir : 7.3.3 (page 192)

Intermede : *Interactions médecin-patient en médecine générale et inégalités sociales de santé – analyses interdisciplinaires*

Période : 2009-11
Financement : Institut de Recherche en Santé Publique
Partenaires : CLLE (CNRS / Université de Toulouse 2), INSERM U558 (Toulouse), LISST (CNRS & UTM), Laboratoire de santé publique et d'épidémiologie (CHU Nantes), LERASS (Université de Toulouse 3) (Toulouse)
Coordinateur : T. Lang (U558)
Rôle : Annotation et exploitation des données

L'objectif du projet INTERMEDE est de comprendre dans quelle mesure les interactions qui se produisent entre le patient et le médecin généraliste dans le cadre de la consultation sont le reflet de certaines inégalités sociales de santé. Il s'appuie sur un corpus de consultations, complété par des entretiens et des questionnaires auprès du patient et du médecin.

Voir : 2.3.3 (page 45), 5.1.1.2 (page 112), 5.2.4 (page 128), 6.4.3 (page 160), 7.2.2 (page 183)

AnnoDis : *Annotation discursive – corpus de référence pour le français et outils d'aide à l'annotation et à l'exploitation*

Période : 2008-10
Financement : Agence Nationale de la Recherche, programme Corpus
Partenaires : CLLE (CNRS / Université de Toulouse 2), IRIT (CNRS / Université de Toulouse 3), GREYC (Université de Caen)
Coordinateur : M.-P. Péry-Woodley (CLLE)
Rôle : Annotation et exploitation des données pour l'approche descendante

Le projet ANNODIS vise la construction d'un corpus de textes annotés au niveau discursif ainsi que le développement d'outils pour l'annotation et l'exploitation de corpus. Les annotations adoptent deux points de vue complémentaires : une perspective ascendante part d'unités de discours minimales pour construire des structures complexes via un jeu de relations de discours ; une perspective descendante aborde le texte dans son entier et se base sur des indices pré-identifiés pour détecter des structures discursives de haut niveau.

Voir : 2.2.3 (page 42), 5.1.2 (page 114), 5.2.2 (page 124, 6.2 (page 144), 6.4.1 (page 159), 7.2.1 (page 182), 7.2.4 (page 185)

Rhécitas : *Rhétorique des citations dans les articles de SHS*

Période : 2008-09
Financement : TGE Adonis (CNRS)
Partenaires : CLLE (CNRS / Université de Toulouse 2), INIST (CNRS, Nancy), IRIT (CNRS / Université de Toulouse 3), Synapse Développement (Toulouse)
Coordinateur : L. Tanguy (CLLE)

Le projet RHECITAS vise à l'identification automatique des fonctions rhétoriques des citations dans les publications en ligne dans le domaine des SHS. Il fait appel pour ce faire à des techniques de TAL pour identifier et caractériser les différents contextes des appels de citation.

Voir : 5.2.3 (page 126), 7.2.3 (page 184)

ARIEL : *Adaptation d'une chaîne de Recherche d'Information sur la base de traitements Linguistiques*

Période : 2004-06
Financement : TCAN (CNRS)
Partenaires : ERSS (CNRS / Université de Toulouse 2), IRIT (CNRS/ Université de Toulouse 3)
Coordinateur : J. Mothe (IRIT) & L. Tanguy (ERSS)

Ce projet vise à étudier différentes techniques et ressources linguistiques pour la définition d'un système de recherche d'information qui s'adapte à l'expression des besoins formulés par l'utilisateur.

Voir : 2.3.2 (page 44), 6.4.2 (page 159)

WESCONVA : *WEb, Suffixation et CONcurrence des déVerbaux d'Action*

Période : 2003-05

Financement : Institut de Linguistique Française

Partenaires : ERSS (CNRS / Université de Toulouse 2), ATILF (CNRS / Université de Nancy 2), SILEX (CNRS / Université de Lille 3)

Coordinateur : G. Dal (SILEX)

Rôle : Mise en place et gestion de la base de données, organisation de l'annotation manuelle, analyse des données.

Ce projet vise l'étude des phénomènes constructionnels concurrents susceptibles de former des déverbaux d'action (*-age, -ment, -tion*) avec une approche quantitative et contextuelle des données. Il se base sur une comparaison entre le lexique attesté dans des corpus de référence (nomenclatures du TLFi et du Robert Électronique) et sur le Web, pour comprendre les raisons motivant une nouvelle création de déverbal, et déterminer les paramètres qui conditionnent le choix de la forme du déverbal.

Voir : 4.4.3.3 (page 98)

YAKWA++ : *Interrogation de corpus étiquetés syntaxiquement*

Période : 2001-04

Financement : Institut de Linguistique Française

Partenaires : ERSS (CNRS / Université de Toulouse 2), Bases, Corpus et Langage (CNRS / Université de Nice)

Coordinateur : D. Bourigault, C. Fabre & L. Tanguy (ERSS)

Ce projet vise à développer une interface d'interrogation de corpus analysés syntaxiquement (en l'occurrence par l'analyseur Syntex) pour un usage linguistique.

Voir : 3.3.1 (page 65)

Étude de l'implantation des termes recommandés

Période : 2001-02

Financement : Délégation Générale à la Langue Française et aux Langues de France)

Coordinateur : D. Bourigault & L. Tanguy (ERSS)

Partenaire : ERSS (CNRS / Université de Toulouse 2)

Ce projet consiste à mesurer l'impact des recommandations faites par la DGLF concernant les termes techniques à employer dans le domaine de l'économie et des finances (par opposition notamment aux termes anglais ou impropres). L'étude s'est faite exclusivement sur un panel de sites Web représentatifs des différents acteurs concernés (institutions, médias, entreprises, écoles).

IDOL : *IRS-Based Document Localisation*

Période : 1997-99

Financement : Communauté européenne, programme INCO-DC)

Partenaires : ISSCO (Suisse), UMIST (Royaume-Uni), EPOS (France), Universal (Tunisie), IME (Liban)

Coordinateur : R. Belhadj Kacem (EPOS)

Rôle : Responsable du module de vérification de la traduction

Ce projet vise à développer une plate-forme d'aide à la traduction pour les langues anglaise, française et arabe. Il propose un ensemble de modules intégrés (mémoire de traduction, gestionnaire terminologique, vérificateur de traduction).

Voir : 1.3.3 (page 36)

DiET : *Diagnosis and Evaluation Tools for Natural Language Processing*

Période : 1996-99

Financement : Communauté européenne (LE 4204)

Partenaires : ISSCO (Suisse), IBM (Allemagne), DFKI (Allemagne), SRI (Royaume-Uni), UCD (Irlande), Aérospatiale (France)

Coordinateur : K. Netter (DFKI)

Rôle : Responsable de l'outil de profilage, permettant d'adapter les bancs de test à une application de TAL spécifique

Ce projet vise à développer une méthode outillée pour l'élaboration de bancs de test permettant l'évaluation de différents systèmes de traitement automatique des langues.

Voir : 1.3.2 (page 35)

Index

Bibliographie

Abeillé, A., L. Clément, et F. Toussenel (2003). Building a treebank for French. In A. Abeillé, éd., *Treebanks*. Kluwer.

Abney, S. (2011). Data-intensive experimental linguistics. *Linguistic Issues in Language Technology*, **6**.

Adam, C. et F. Morlane-Hondère (2009). Détection de la cohésion lexicale par voisinage distributionnel : application à la segmentation thématique. In *Actes de la conférence RECITAL'09*. Senlis, France.

Adam, C., P. Muller, et C. Fabre (2010). Une évaluation de l'impact des types de textes sur la tâche de segmentation thématique. In *Actes de TALN*.

Adda, G., J. Mariani, P. Paroubek, M. Rajman, et J. Lecomte (1999). L'action grace d'évaluation de l'assignation de parties du discours pour le français. *Cahiers Langues*, **2**(2) :119–129.

Armstrong, S., P. Bouillon, et G. Robert (1995). Tatoo tagger overview. Rapport technique, ISSCO, `http://isscowww.unige.ch/staff/robert/tatoo/tagger.html`.

Atkins, B. S. et M. Rundell (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

Aussenac-Gilles, N. et A. Condamines (2009). Variation syntaxique et contextuelle dans la mise au point de patrons de relations sémantiques. In J.-L. Minel, éd., *Filtrage sémantique*, pp. 115–149. Hermès/Lavoisier.

Baayen, H., H. van Halteren, et F. Tweedie (1996). Outside the cave of shadows : using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, **11**(3) :121–31.

Baayen, R. H. (2001). *Word frequency distributions*. Kluwer.

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baayen, R. H. et A. Neijt (1997). Productivity in context : a case study of a dutch suffix. *Linguistics*, **35** :565–587.

Baroni, M. et S. Bernardini (2004). Bootcat : bootstrapping corpora and terms from the web. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*. Lisbon.

Baroni, M., S. Bernardini, A. Ferraresi, et E. Zanchetta (2009). The wacky wide web : A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, **43**(3) :209–226.

Baroni, M., F. Chantree, A. Kilgarriff, et S. Sharoff (2008). Cleaneval : a Competition for Cleaning Web Pages. In *Proceedings of LREC*. Marrakech.

Baroni, M. et A. Lenci (2010). Distributional memory : A general framework for corpus-based semantics. *Computational Linguistics*, **36**(4) :673–721.

Beaudouin, V., S. Fleury, B. Habert, G. Illiouz, C. Licoppe, et M. Pasquier (2001). Typweb : Décrire la toile pour mieux comprendre les parcours. In *CIUST'01 : Colloque International sur les Usages et les Services de Télécommunications*. Paris.

Benzécri, J.-P. (1982). *L'analyse des données*. Dunod.

Berger, A. L., S. A. Della Pietra, et V. J. Della Pietra (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, **22**.

Bernard, P., J. Dendien, J. Lecomte, et J.-M. Pierrel (2002). Les ressources de l'ATILF pour l'analyse lexicale et textuelle : TLFi, Frantext et le logiciel Stella. In A. Morin et P. Sébillot, éds., *6es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2002)*, pp. 137–148. Saint-Malo.

Bertin, J. (1970). La graphique. *Communications*, **15** :169–185.

Beust, P. (1998). *Contribution à un modèle interactionniste du sens*. Thèse de doctorat, Université de Caen.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D. (1995). On the role of computational, statistical and interpretive techniques in multi-dimensional analysis of register variation. *Text*, **15**(3) :314–370.

Bilhaut, F. et A. Widlöcher (2006). Linguastream : An integrated environment for computational linguistics experimentation. In *Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics (EACL)*, pp. 95–98. Trento, Italy.

Bird, S., Y. Chen, S. Davidson, H. Lee, et Y. Zheng (2006). Designing and evaluating an XPath dialect for linguistic queries. In *Proceedings of the 22nd International Conference on Data Engineering, ICDE'06*.

Bommier-Pincemin, B. (1999). *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*. Thèse de doctorat, Université de Paris IV.

Bouffier, A. (2009). Une approche textuelle pour l'analyse de textes de recommandations médicales. *TAL*, **50**(1) :35–59.

Bouillon, P., R. Baud, G. Robert, et P. Ruch (2000). Indexing by statistical tagging. In *Actes des 5esjournées d'analyse statistique des données textuelles (JADT)*.

Bouillon, P., S. Lehmann, I. Lewin, D. Milward, et L. Tanguy (1999). Discourse data in DiET. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora (LINC)*. Bergen, Norway.

Bourigault, D. (1995). Lexter, a terminology extraction software for knowledge acquisition from texts. In *Proceedings of the 9th Knowledge Acquisition for Knowledge Based System Workshop (KAW'95)*. Banff, Canada.

Bourigault, D. (2007). *Un analyseur syntaxique opérationnel : SYNTEX*. Mémoire d'habilitation à diriger des recherches, Université de Toulouse 2.

Bourigault, D., C. Fabre, C. Frérot, M.-P. Jacques, et S. Ozdowska (2005). Syntex, analyseur syntaxique de corpus. In *Actes de TALN*. Dourdan.

Brants, T. et A. Franz (2006). Web 1t 5-gram corpus version 1.1. Linguistic Data Consortium.

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing*. Trento, Italy.

Carmel, D. et E. Yom-Tov (2010). *Estimating the Query Difficulty for Information Retrieval*. Morgan and Claypool.

Champely, S. (2004). *Statistique appliquée au sport*. de Boeck.

Chrisment, C., T. Dkaki, J. Mothe, S. Poulain, et L. Tanguy (2005). Recherche d'information : analyse de différents systèmes réalisant la même tâche. *Ingénierie des systèmes d'information*, **10**(1) :33–58.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Proceedings 3rd Conference on Computational Lexicography and Text Research (COMPLEX '94)*. Budapest, Hungary.

Church, K. (2011). A pendulum swung too far. *Linguistic Issues in Language Technology*, **6**.

Church, K. W. et R. L. Mercer (1993). Introduction to the special issue on using large corpora. *Computational Linguistics*, **19**(1).

Cohen, W. W. (1995). Fast effective rule induction. In *Twelfth International Conference on Machine Learning*, pp. 115–123.

Condamines, A. (2005a). Anaphore nominale infidèle et hyperonymie : le rôle du genre textuel. *Revue de Sémantique et Pragmatique*, **18** :23–42.

Condamines, A. (2005b). Sémantique et corpus : quelles rencontres possibles ? In A. Condamines, éd., *Sémantique et Corpus*, pp. 17–38. Hermès.

Condamines, A., C. Fabre, et M.-P. P. Woodley, éds. (1999). *Actes de l'Atelier Corpus et TAL : Pour une réflexion méthodologique*. TALN99, Cargèse.

Cori, M. et J. Léon (2002). La constitution du tal : Étude historique des dénominations et des concepts. *TAL*, **43**(3) :21–55.

Crabbé, B. et M.-H. Candito (2008). Expériences d'analyse syntaxique du français. In *Actes de TALN*. Avignon.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, et W. Peters (2011). *Text Processing with GATE (Version 6)*. University of Sheffield Department of Computer Science.

Dal, G., S. Lignon, F. Namer, et L. Tanguy (2004). Toile contre dictionnaires : analyse morphologique en corpus de noms déverbaux concurrents. In *Colloque International sur "Les noms déverbaux"*. Villeneuve d'Ascq.

Dal, G. et F. Namer (2010). Les noms en -ance/-ence du français : quel(s) patron(s) constructionnel(s) ? In *Actes du 2e Congrès Mondial de Linguistique Française*, pp. 893–907. Nouvelle Orléans, Etats-Unis.

Daoust, F. (1996). Sato 4. Manuel de référence, Centre ATO, UQAM, Montréal.

Daumé, H. C. (2006). *Practical Structured Learning Techniques for Natural Language Processing*. Thèse de doctorat, University of Southern California.

De Schryver, G.-M. (2002). Web for / as corpus : a perspective for the african languages. *Nordic Journal of African Studies*, **11**(2) :266–282.

Duclaye, A., , F. Yvon, et O. Collin (2002). Using the web as a linguistic resource for learning reformulations. In *Proceedings of the third international conference on language resources and evaluation (LREC'02*. Citeseer.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, **19**(1) :61–74.

Fabre, C. (2010). *Unités syntaxiques et sémantiques entre les mots : apports mutuels de la linguistique et du traitement automatique des langues*. Habilitation à diriger des recherches, Université de Toulouse 2.

Fabre, C. et D. Bourigault (2006). Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. In *Actes de TALN*, pp. 121–129. Leuven, Belgique.

Fairon, C., K. Macé, et H. Naets (2008). Glossanet 2 : A linguistic search engine for rss-based corpora. In *Proceedings of the 4th web as corpus workshop (WAC-4)*, pp. 34–39.

Falaise, A., A. Tutin, et O. Kraif (2011). Exploration d'un corpus arboré pour non spécialistes par des requêtes guidées et des requêtes sémantiques. In *Actes de TALN*.

Fang, X., C. Jacquemin, F. Vernier, et B. Luo (2009). A survey of 3d document corpus visualization. *Information Technology Journal*, **8**(1) :1–15.

Fekete, J.-D., J. van Wijk, J. Stasko, et C. North (2008). The value of information visualization. In A. Kerren, J. Stasko, J.-D. Fekete, et C. North, éds., *Information Visualization – Human-Centered Issues and Perspectives*. Springer.

Fletcher, W. (2006). Concordancing the web : promise and problems, tools and techniques. *Language and Computers*, **59**(1) :25–45.

Friedman, J. H. (1997). Data mining and statistics : What's the connection ? In *Proceedings of Computer Science and Statistics : the 29th Symposium on the Interface*.

Friedman, J. H. (2006). Recent advances in predictive (machine) learning. *Journal of Classification*, **2**(23) :175–197.

Frérot, C., D. Bourigault, et C. Fabre (2003). Marier apprentissage endogène et ressources exogènes dans un analyseur syntaxique de corpus. le cas du rattachement verbal à distance de la préposition « de ». *TAL*, **44**(3).

Gala, N. (2003). Une méthode non supervisée d'apprentissage sur le web pour la résolution d'ambiguïtés structurelles liées au rattachement prépositionnel. In *Actes de TALN*. Batz-sur-Mer.

Gale, W. A. et K. W. Church (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, **19**(1) :75–102.

Gamon, M. (2004). Sentiment classification on customer feedback data : Noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Ghiglione, R., A. Landré, M. Bromberg, et P. Molette (1998). *L'analyse automatique des contenus*. Dunod.

Grefenstette, G. (1998). The world wide web as a resource for example-based machine translation tasks. In *Proceedings of the ASLIB Conference on Translating and the Computer*. London.

Greimas, A. J. (1966). *Sémantique structurale : recherche et méthode*. Larousse.

Gries, S. T. (2005). Null-hypothesis significance testing on word frequencies : a follow-up on kigarriff. *Corpus linguistics and linguistics theory*, **1**(2) :277–294.

Gries, S. T. (2009). *Quantitative corpus linguistics with R : a practical introduction*. Routledge.

Génolini, J.-P., R. Roca, C. Rolland, et M. Membrado (2011). « L'éducation » du patient en médecine générale : une activité périphérique ou spécifique de la relation de soin ? *Revue Sciences Sociales et Santé*, **29**(3) :81–122.

Habert, B. (2004). Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs. *Revue française de linguistique appliquée*, **IX**(1).

Habert, B. (2006). Portrait de linguiste(s) à l'instrument. In C. Guillot, S. Heiden, et S. Prévost, éds., *À la quête du sens : études littéraires, historiques et linguistiques en hommage à Christiane Marchello-Nizia*, pp. 124–132. ENS Editions.

Habert, B., A. Nazarenko, et A. Salem (1997). *Les linguistiques de corpus*. Armand Colin.

Hajičová, E. (2011). Computational linguistics without linguistics ? view from prague. *Linguistic Issues in Language Technology*, **6**.

Hall, D., D. Jurafsky, et C. D. Manning (2008). Studying the history of ideas using topic models. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 363–371.

Hathout, N. (2000). Morphological pairing based on the network model. In *Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX 2000)*, pp. 35–38. Pyrgos, Grèce.

Hathout, N., F. Montermini, et L. Tanguy (2008). Extensive data for morphology : using the World Wide Web. *Journal of French Language Studies*, **18**(1) :67–85.

Hathout, N., F. Namer, et G. Dal (2002). An experimental constructional database : The mortal project. In P. Boucher, éd., *Many Morphologies*. Cascadilla, Somerville, Mass.

Hathout, N., M. Plénat, et L. Tanguy (2004). Enquête sur les dérivés en -able. *Cahiers de Grammaire*, **28** :49–90.

Hathout, N., F. Sajous, et L. Tanguy (2009). Looking for French deverbal nouns in an evolving Web (a short history of WAC). In *Proceedings of the Fifth Workshop on Web As Corpus (WAC)*, pp. 37–44. San-Sebastian, Spain.

Hathout, N. et L. Tanguy (2002). Webaffix : a tool for finding and validating morphological links on the WWW. In *Proceedings of LREC*. Las Palmas, Spain.

Hathout, N. et L. Tanguy (2005). WEBAFFIX : une boîte à outils d'acquisition lexicale à partir du Web. *Revue Québécoise de Linguistique*, **32**(1) :61–84.

Have, P. t. (1989). The consultation as a genre. In B. Torode, éd., *Text and talk as social practice*, pp. 115–133. Foris Publications.

Hearst, M. (1997). Texttiling : segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, **23**(1) :33–64.

Heiden, S., J.-P. Magué, et B. Pincemin (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In *Acetes de la conférence JADT*.

Henestroza Anguiano, E. et M. Candito (2011). Resolving difficult syntactic attachments with parse correction. In *Proceedings of EMNLP'2011*. Edimburg.

Henry, N. et J.-D. Fekete (2008). Représentations visuelles alternatives pour les réseaux sociaux. *Réseaux*, **152**(6) :59–92.

Hermann, E., S. Leblois, M. Mazeau, D. Bourigault, C. Fabre, S. Travadel, P. Durgeat, et D. Nouvel (2008). Outils de traitement automatique des langues appliqués aux comptes rendus d'incidents et d'accidents. In *actes du 16^e Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement*.

Ho-Dac, L.-M., M.-P. Péry-Woodley, et L. Tanguy (2010). Anatomie des structures énumératives. In *Actes de TALN'2010*. Montréal.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th Annual International ACM SIGIR Conference*, pp. 329–338. New York, USA.

Hundt, M., N. Nesselhauf, et C. Biewer, éds. (2007). *Corpus linguistics and the Web*. Rodopi, Amsterdam.

Hurter, C. (2010). *Caractérisation de Visualisations et Exploration Interactive de Grandes Quantités de Données Multidimensionnelles*. Thèse de doctorat, Université de Toulouse.

Ide, N. et C. Brew (2000). Requirements, tools and architectures for annotated corpora. In *Proceedings of the Workshop on Data Architectures and Software Support for Large Corpora, LREC*, pp. 1–5. Athens, Grèce.

Jacquemin, C. (1997). *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. Mémoire d'habilitation à diriger des recherches, Université de Nantes.

Jacquemin, C. et C. Bush (2000). Fouille du web pour la collecte d'entités nommées. In *Actes de TALN*. EPFL, Lausanne.

Jacquemin, C., H. Folch, K. Garcia, et S. Nugier (2005). Visualisation interactive d'espaces documentaires. *Revue Information - Interaction - Intelligence*, **5**(1).

Jacques, M.-P. et N. Aussenac-Gilles (2006). Variabilité des performances des outils de tal et genre textuel. *Traitement automatique des langues*, **47**(1) :11–32.

Jelinek, F. (2005). Some of my best friends are linguists. *Language Resources and Evaluation*, **1**(39) :25–34.

Johnson, B. et B. Shneiderman (1991). Treemaps : a space-filling approach to the visualization of hierarchical information structures. In *Proceedings of the 2nd International IEEE Visualization Conference*, pp. 284–291. Sand Diego.

Johnson, M. (2011). How relevant is linguistics to computational linguistics ? *Linguistic Issues in Language Technology*, **6**.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, **1**(3) :233–334.

Jäverlin, K. (2009). Explaining user performance in information retrieval : Challenges to ir evaluation. In *Proceedings of the 2nd International Conference on the Theory of Information Retrieval*, pp. 289–296.

Kanellos, I., J. Le Dû, F. Legras, et L. Tanguy (1999). Assistance informatique à l'interprétation des données en cartographie linguistique : Informatisation anthropocentrée du Nouvel Atlas Linguistique de la Basse-Bretagne. *Géolinguistique*, **8** :181–196.

Kay, M. (2011). Zipf's law and l'arbitraire du signe. *Linguistic Issues in Language Technology*, **6**.

Kehoe, A. et A. Renouf (2002). Webcorp : applying the web to linguistics and linguistics to the web. In *Proceedings of the WWW 2002 Conference*. Honolulu.

Keim, D., G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, et G. Melançon (2008). Visual Analytics : Definition, Process, and Challenges. In A. Kerren, J. T. Stasko, J.-D. Fekete, et C. North, éds., *Information Visualization – Human-Centered Issues and Perspectives*, LNCS State-of-the-Art Survey, pp. 154–175. Springer.

Keller, F. et M. Lapata (2003). Using the web to obtain frequencies for unseen bigrams. *Computational linguistics*, **29**(3) :459–484.

Kelling, C. (2003). The role of agentivity for suffix selection. In *Proceedings of the Third Mediterranean Meeting on Morphology*.

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, **2**(1).

Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics*, **33**(1).

Kilgarriff, A. et G. Grefenstette (2003). Introduction to the special issue on web as corpus. *Computational Linguistics*, **29**(3) :333–347.

Kim, Y.-M., P. Bellot, E. Faath, et M. Dacos (2011). Automatic annotation of bibliographical references in digital humanities books, articles and blogs. In G. Kazai, C. Eickhoff, et P. Brusilovsky, éds., *BooksOnline*, pp. 41–48. ACM.

Klavans, J. L. et P. Resnik, éds. (1996). *The balancing act : combining symbolic and statistical approaches to language*. MIT Press.

Klein, D. et C. D. Manning (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423–430.

Klimt, B. et Y. Yang (2004). Introducing the enron corpus. In *First Conference on Email and Anti-Spam (CEAS)*. Mountain View, CA.

Koppel, M., J. Schler, et S. Argamon (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, **60**(1) :9–26.

König, E., W. Lezius, et H. Voormann (2003). *TIGERSearch user's manual*. IMS, University of Stuttgart.

Labbé, C. et D. Labbé (2001). Inter-textual distance and authorship attribution : Corneille and molière. *Journal of Quantitative Linguistics*, **8**(3) :213–231.

Lafon, P. (1981). Statistiques des localisations des formes d'un texte. *Mots*, **2**(2) :157–188.

Lai, C. et S. Bird (2004). Querying and updating treebanks : a critical survey and requirements analysis. In *Proceedings of the Australasian Language Technology Workshop*, pp. 139–146.

Laignelet, M. (2009). *Analyse discursive pour le repérage automatique de segments obsolescents dans des documents encyclopédiques.* Thèse de doctorat, Université de Toulouse 2.

Laignelet, M., M.-P. Péry-Woodley, et L. Tanguy (2010). Découverte de configurations de traits textuels pour la caractérisation des segments d'obsolescence. *Document Numérique*, **13**(3) :41–68.

Laignelet, M. et F. Rioult (2010). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. *TAL*, **51**(1).

Lamiroy, B. et M. Charolles (2010). Les clitiques accusatifs versus datifs dans les constructions causatives en faire. In *actes du 2ème Congrès Mondial de Linguistique Française*.

Le Dû, J. (2001). *Nouvel Atlas Linguistique de la Basse-Bretagne*. CRBC.

Lebart, L., M. Piron, et A. Morineau (2006). *Statistique Exploratoire Multidimensionnelle*. Dunod.

Lebart, L. et A. Salem (1994). *Statistique textuelle*. Dunod.

Lecomte, J. (1998). Le catégoriseur brill14-jl5 / winbrill-0.3. Rapport technique, INaLF, `http://www.atilf.fr/winbrill`.

Lehmann, S., S. Oepen, S. Regnier-Prost, K. Netter, V. Lux, J. Klein, K. Falkedal, F. Fouvry, D. Estival, E. Dauphin, H. Compagnion, J. Baur, L. Balkan, et D. Arnold (1996). Tsnlp : Test suites for natural language processing. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*, COLING '96, pp. 711–716.

Leroy, S. (2004). Extraire sur patrons : allers et retours entre analyse linguistique et repérage automatique. *Revue Française de linguistique appliquée*, **9**(1) :25–43.

Levin, L. (2011). Variety, idiosyncracy, and complexity in language and language technologies. *Linguistic Issues in Language Technology*, **6**.

Li, W., J. Han, et J. Pei (2001). Cmar : Accurate and efficient classification based on multiple class-association rules. In *International Conference on Data Mining (ICDM'01)*. San Jose.

Lin, D., K. Church, H. Ji, S. Sekine, D. Yarowsky, S. Bergsma, K. Patil, E. Pitler, R. Lathbury, V. Rao, *et al.* (2010). New tools for web-scale n-grams. In *Proceedings of LREC*.

Longrée, D. et S. Mellet (2010). Analysis of textual data, some topological methods for studying text structure indicators : the case of latin historic narratives. In *Proceedings of the International Workshop on Multidisciplinary Approaches to Discourse (MAD)*. Moissac.

Lüdeling, A., S. Evert, et M. Baroni (2007). Using web data for linguistic purposes. In Hundt *et al.* (2007).

Mandl, T. et C. Womser-Hacker (2002). Linguistic and statistical analysis of the clef topics. In *Proceedings of the CLEF 2002 Workshop*.

Manning, C. et H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Marcus, M. P., M. A. Marcinkiewicz, et B. Santorini (1993). Building a large annotated corpus of English : the Penn Treebank. *Computational Linguistics*, **19**(2).

Martin, F. (2008). The semantics of eventive suffixes in french. In F. Schäfer, éd., *'SinSpec', Working Papers of the SFB 732*. University of Stuttgart.

Martin, F. (2012, à paraître). Stage level and individual level readings of quality nouns. In N. Hathout, F. Montermini, et J. Tseng, éds., *Actes des 7^eDécembrettes*. Lincom Europa.

Mazza, R. (2009). *Introduction to information visualization*. Springer.

Mellet, S. et J.-P. Barthélemy (2009). La topologie textuelle : légitimation d'une notion émergente. *Lexicometrica*, **7**.

Milard, B. (2011). *Activités scientifiques, textes et réseaux sociaux. Dynamiques relationnelles à travers les citations, publications et bases de données de la recherche scientifique*. Mémoire d'habilitation à diriger des recherches, Université de Toulouse 2.

Missire, R. (2005). Rythmes sémantiques et temporalité des parcours interprétatifs. Texte en ligne sur le site Texto (`http://www.revue-texto.net/`).

Moreau, F., V. Claveau, et P. Sébillot (2007). Intégrer plus de connaissances linguisitiques en recherche d'information peut-il augmenter les performances des systèmes ? In *4^econférence en recherche d'informations et applications, CORIA'07*, pp. 223–238. Saint-Étienne.

Mothe, J. et L. Tanguy (2005). Linguistic features to predict query difficulty. In *proceedings of the ACM-SIGIR workshop on Predicting query difficulty - methods and applications*, pp. 7–10. Salvador de Bahia, Brazil.

Mothe, J. et L. Tanguy (2007). Linguistic Analysis of Users' Queries : towards an adaptive Information Retrieval System. In *proceedings of the IEEE International Conference on Signal-Image Technology & Internet-Based Systems*. Shangai.

Mourlhon-Dallies, F., F. Rakotonoelina, et S. Reboul-Touré (2004). Les discours de l'internet : quels enjeux pour la recherche ? *Les Carnets du Cediscor*, **8**.

Namer, F. (2003). Valider les unités morphologiques par le web. In *Silexicale3, Actes du 3e Forum de Morphologie*.

Pauw, G. D. et P. W. Wagacha (2007). Bootstrapping morphological analysis of gĩkũyũ using maximum entropy learning. In *Proceedings of the eighth INTERSPEECH conference*.

Pédauque, R. T. (2003). Document : forme, signe et médium, les re-formulations du numérique. `http://archivesic.ccsd.cnrs.fr/sic_00000511`.

Péry-Woodley, M.-P., N. Asher, P. Enjalbert, F. Benamara, M. Bras, C. Fabre, S. Ferrari, L.-M. Ho Dac, A. Le Draoulec, Y. Mathet, P. Muller, L. Prévot, J. Rebeyrolle, L. Tanguy, M. Vergez Couret, L. Vieu, et A. Widlöcher (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. In *Actes de TALN*, p. 52. Senlis.

Plénat, M., S. Lignon, N. Serna, et L. Tanguy (2002). La conjecture de Pichon. *Corpus*, **1** :105–150.

Plénat, M. (1997). Analyse morphophonologique d'un corpus d'adjectifs dérivés en *-esque*. *Journal of French Language Studies*, **7** :163–179.

Plénat, M. et G. Boyé (2012, à paraître). Le choix des thèmes dans les dérivés désadjectivaux en français. In B. Tranel, éd., *Understanding Allomorphy. Perspectives from Optimality Theory*. Equinox.

Poibeau, T. (2003). *Extraction automatique d'information*. Hermès.

Péry-Woodley, M.-P. (1995). Quels corpus pour quels traitements automatiques ? *TAL*, **36**(1-2) :213–232.

Péry-Woodley, M.-P. (1998). Signalling in written text : a corpus-based approach. In M. Stede, L. Wanner, et E. Hovy, éds., *Discourse Relations and Discourse Markers (COLING'98 Workshop)*, pp. 79–85.

Péry-Woodley, M.-P. (2005). Discours, corpus, traitements automatiques. In A. Condamines, éd., *Sémantique et Corpus*, pp. 177–205. Hermès.

Quinlan, R. (1993). *C4.5 : Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Rastier, F. (1987). *Sémantique Interprétative*. Presses Universitaires de France.

Rastier, F. (1989). *Sens et Textualité*. Hachette.

Rastier, F. (1991). *Sémantique et recherches cognitives*. Presses Universitaires de France.

Rastier, F. (2005). Enjeux épistémologiques de la linguistique de corpus. In G. Williams, éd., *La Linguistique de corpus*, pp. 31–46. Presses Universitaires de Rennes.

Rastier, F., M. Cavazza, et A. Abeillé (1994). *Sémantique pour l'analyse : de la linguistique à l'informatique*. Masson.

Rebeyrolle, J. (2000). *Forme et fonction de la définition en discours*. Thèse de doctorat, Université de Toulouse 2.

Rebeyrolle, J., D. Bourigault, C. Fabre, A. Josselin Leray, et L. Tanguy (2007). Un laboratoire d'observation de l'usage du vocabulaire recommandé par les instances officielles françaises. In *Colloque Presciptions En Langue*. Paris.

Rebeyrolle, J. et L. Tanguy (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire*, **25** :153–174.

Reinert, M. (1990). Alceste, une méthodologie d'analyse des données textuelles et une application : Aurelia de gerard de nerval. *Bulletin de méthodologie sociologique*, **26**(1) :24–54.

Resnik, P. (1999). Mining the web for bilingual text. In *37th Meeting of ACL*, pp. 527–534. Maryland, USA.

Resnik, P., A. Elkiss, E. Lau, et H. Taylor (2005). The web in theoretical linguistics research : Two case studies using the linguist's search engine. In *proceedings of the 31st Meeting of the Berkeley Linguistics Society*, pp. 265–276.

Roché, M. (2008). Structuration du lexique et principe d'économie : le cas des ethniques. In *actes du premier Congrès Mondial de Linguistique Française*.

Rohde, D. (2005). *Tgrep2 User Manual*. `http://tedlab.mit.edu/~dr/Tgrep2/tgrep2.pdf`.

Rundell, M. (2000). The biggest corpus of all. *Humanising Language Teaching*, **2**(3).

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, **4**(4) :248–375.

Santini, M. (2007). Characterizing genres of web pages : Genre hybridism and individualization. In *Proceedings of the 40th Hawaii International Conference on System Sciences*.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pp. 44–49. Manchester, UK.

Schmouchkovitch, M., L. Tanguy, et I. Kanellos (1998). Code et systèmes de significations dans un cas de délire interprétatif. *Annales Médico-Psychologiques*, **156**(6) :375–386.

Séguéla, P. et N. Aussenac-Gilles (1999). Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In *Actes de la conférence IC (Ingénierie des Connaissances)*, pp. 79–88.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In Baroni et Bernardini, éds., *Wacky ! Working Papers on the Web as Corpus*. GEDIT.

Shneiderman, B. (1996). The eyes have it : A task by data type taxonomy for information visualizations. In I. C. S. Press, éd., *Proceedings of the IEEE Symposium on Visual Languages*, pp. 336–343.

Silberztein, M. (1993). *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*. Masson.

Sinclair, J. (2004). Corpus and text : basic principles. In M. Wynne, éd., *Developing Linguistic Corpora. A guide to Good Practice*, pp. 1–16. Oxbow.

Smith, N. A. (2011). *Linguistic Structure Prediction*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

Smucker, M. D., J. Allan, et B. Carterette (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, pp. 623–632. New York, USA.

Sparck-Jones, K. (1999). What is the role of NLP in text retrieval ? In T. Strzalkowski, éd., *Natural Language Information Retrieval*, Speech and Language Technology series. Kluwer, Dordrecht.

Steedman, M. (2011). Romantics and revolutionaries. *Linguistic Issues in Language Technology*, **6**.

Swales, J. M. (1990). *Genre Analysis*. Cambridge Uniersity Press.

Tang, M., X. Luo, et S. Roukos (2002). Active learning for statistical natural language parsing. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 120–127. Philadelphia.

Tanguy, L. (1997). *Traitement automatique de la langue naturelle et Interprétation : Contribution à l'élaboration d'un modèle informatique de la Sémantique Interprétative*. Thèse de doctorat, Université de Rennes 1.

Tanguy, L., S. Armstrong, P. Boullon, et S. Lehmann (1999a). DiET : Diagnostic et evaluation des systèmes de traitement de la langue naturelle. *Cahiers Langues*, **2** :140–150.

Tanguy, L., S. Armstrong, et D. Walker (1999b). Isotopies sémantiques et vérification de traduction. In *Actes de TALN*. Cargèse.

Tanguy, L., C. Fabre, L.-M. Ho-Dac, et J. Rebeyrolle (2011a). Caractérisation des échanges entre patients et médecins : approche outillée d'un corpus de consultations médicales. *Corpus*, **10** :137–154.

Tanguy, L. et N. Hathout (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web. In *Actes de TALN'2002*, p. 254. Nancy France.

Tanguy, L. et N. Hathout (2007). *Perl pour les linguistes : programmes en Perl pour exploiter les données langagières*. TIC et Sciences Cognitives. Hermès sciences publications.

Tanguy, L., F. Lalleman, C. François, P. Muller, et P. Séguéla (2009). RHECITAS : citation analysis of French humanities articles. In *Proceedings of Corpus Linguistics*. Liverpool, UK.

Tanguy, L. et N. Tulechki (2009). Sentence Complexity in French : a Corpus-Based Approach. In *Proceedings of of the International Conference on Recent Advances in Intelligent Information Systems (IIS)*, pp. 131–145. Krakow, Poland.

Tanguy, L., A. Urieli, B. Calderone, N. Hathout, et F. Sajous (2011b). A multitude of linguistically-rich features for authorship attribution. In *Notebook for PAN at CLEF 2011*. Amsterdam.

Tellier, I. (2009). Préface au numéro spécial sur l'apprentissage automatique pour le TAL. *TAL*, **50**(3).

Teufel, S., A. Siddharthan, et D. Tidhar (2006). Automatic classification of citation function. In *Proceedings of EMNLP 6*, pp. 103–110.

Thlivitis, T. (1998). *Sémantique Interprétative Intertextuelle : assistance informatique anthropocentrée à la compréhension des textes*. Thèse de doctorat, Université de Rennes 1.

Thom, R. (1993). *Prédire n'est pas expliquer (entretiens avec Emile Noël)*. Flammarion.

Todorovski, L., B. Cestnik, M. Kline, N. Lavrač, et S. Džeroski (2002). Qualitative clustering of short time-series : A case study of firms reputation data. In M. Bohanec, B. Kavšek, N. Lavrač, et D. Mladenic, éds., *IDDM02*, pp. 141–149. Helsinki University Printing House.

Tomanek, K. et F. Olsson (2009). A web survey on the use of active learning to support annotation of text data. In *Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing*, pp. 45–48. Boulder, Colorado.

Tulechki, N. (2011). Des outils de tal en support aux experts de sûreté industrielle pour l'exploitation de bases de données de retour d'expérience. In *Actes de Recital*. Montpellier.

Tutin, A., F. Grossmann, A. Falaise, et O. Kraif (2009). Autour du projet Scientext : étude des marques linguistiques du positionnement de l'auteur dans les écrits scientifiques. In *Actes des 6esjournées de linguistique de corpus*.

Valette, M. (2009). *Approche textuelle du lexique*. mémoire d'habilitation à diriger des recherches, INALCO.

Valette, M. (2010). Méthodes pour la veille lexicale. In *Actes de la journée d'étude sur Le dictionnaire électronique : quelles perspectives pour les sciences humaines et sociales ?*, pp. 15–29.

Valette, M. et N. Grabar (2004). Caractérisation de textes à contenu idéologique : statistique textuelle ou extraction se syntagme ? l'exemple du projet PRINCIP. In *Actes des 7èmes Journées Internationale d'Analyse statistique des Données Textuelles (JADT)*, pp. 1106–1116. UCL-Presses Universitaires de Louvain, Louvain-la-Neuve, Belgique.

Valette, M. et F. Rastier (2006). Prévenir le racisme et la xénophobie – propositions de linguistes. *Les langues modernes*, **2** :68–77.

Vergely, P. (2004). *Analyse linguistique de l'expression du dysfonctionnement technique : le cas des échanges entre chefs de salle et maintenance opérationnelle dans la Navigation Aérienne*. Thèse de doctorat, Université de Toulouse 2.

Vergely, P., A. Condamines, C. Fabre, A. Josselin-Leray, J. Rebeyrolle, et L. Tanguy (2009). Analyse linguistique des interactions patient/médecin. In *Actes du colloque Actes éducatifs et de soins*. Nice.

Visetti, Y.-M. (1991). Des systèmes experts aux systèmes à base de connaissances : à la recherche d'un nouveau schéma régulateur. *Intellectica*, **12** :221–279.

Wang, K., C. Thrasher, E. Viegas, X. Li, et B. june Hsu (2010). An overview of Microsoft Web N-gram corpus and applications. In *Proceedings of the NAACL-HLT demonstration session)*, pp. 45–48.

Widdows, D. (2004). *Geometry and Meaning*. CSLI publications.

Widlöcher, A. et F. Bilhaut (2005). La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus. In *Actes de TALN*. Dourdan.

Wildlöcher, A. et Y. Mathet (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. In *Actes de TALN*. Senlis.

Witten, I. H. (2005). Text mining. In M. Singh, éd., *The Practical Handbook of Internet Computing*. Chapman & Hall.

Witten, I. H. et E. Frank (2005). *Data Mining : practical learning tools and techniques*. Elsevier.

Wooldridge, R. (2004). Le web comme corpus d'usages linguistiques. *Cahiers de lexicologie*, **85** :209–225.

Xu, J. et W. B. Croft (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transaction on Information Systems*, **16**(1) :61–81.

Yvon, F. (2002). Classification approaches for linguistic analysis. In *Matemáticas y tratamiento de corpus*, pp. 127–144. Fondación San Millán de la Cogola, Logro no, Spain.

Yvon, F. (2006). *Des apprentis pour le traitement automatique des langues*. Mémoire d'habilitation à diriger des recherches, Université Pierre et Marie Curie, Paris.